


## Article

# Variational Message Passing and Local Constraint Manipulation in Factor Graphs

İsmail Şenöz <sup>1,\*</sup> , Thijs van de Laar <sup>1</sup>, Dmitry Bagaev <sup>1</sup> and Bert de Vries <sup>1,2</sup>

<sup>1</sup> Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands; T.W.v.d.Laar@tue.nl (T.v.d.L.); d.v.bagaev@tue.nl (D.B.); bert.de.vries@tue.nl (B.d.V.)  
<sup>2</sup> GN Hearing, JF Kennedylaan 2, 5612 AB Eindhoven, The Netherlands  
\* Correspondence: i.senoz@tue.nl

**Abstract:** Accurate evaluation of Bayesian model evidence for a given data set is a fundamental problem in model development. Since evidence evaluations are usually intractable, in practice variational free energy (VFE) minimization provides an attractive alternative, as the VFE is an upper bound on negative model log-evidence (NLE). In order to improve tractability of the VFE, it is common to manipulate the constraints in the search space for the posterior distribution of the latent variables. Unfortunately, constraint manipulation may also lead to a less accurate estimate of the NLE. Thus, constraint manipulation implies an engineering trade-off between tractability and accuracy of model evidence estimation. In this paper, we develop a unifying account of constraint manipulation for variational inference in models that can be represented by a (Forney-style) factor graph, for which we identify the Bethe Free Energy as an approximation to the VFE. We derive well-known message passing algorithms from first principles, as the result of minimizing the constrained Bethe Free Energy (BFE). The proposed method supports evaluation of the BFE in factor graphs for model scoring and development of new message passing-based inference algorithms that potentially improve evidence estimation accuracy.



**Citation:** Şenöz, İ.; van de Laar, T.; Bagaev, D.; de Vries, B. Variational Message Passing and Local Constraint Manipulation in Factor Graphs. *Entropy* **2021**, *23*, 807. <https://doi.org/10.3390/e23070807>

Academic Editor: Pierre Alquier

Received: 19 May 2021  
Accepted: 22 June 2021  
Published: 24 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** Bayesian inference; Bethe free energy; factor graphs; message passing; variational free energy; variational inference; variational message passing

## 1. Introduction

Building models from data is at the core of both science and engineering applications. The search for good models requires a performance measure that scores how well a particular model  $m$  captures the hidden patterns in a data set  $D$ . In a Bayesian framework, that measure is the *Bayesian evidence*  $p(D|m)$ , i.e., the probability that model  $m$  would generate  $D$  if we were to draw data from  $m$ . The art of modeling is then the iterative process of proposing new model specifications, evaluating the evidence for each model and retaining the model with the most evidence [1].

Unfortunately, Bayesian evidence is intractable for most interesting models. A popular solution to evidence evaluation is provided by *variational* inference, which describes the process of Bayesian evidence evaluation as a (free energy) minimization process, since the variational free energy (VFE) is a tractable upper bound on Bayesian (negative log-)evidence [2]. In practice, the model development process then consists of proposing various candidate models, minimizing VFE for each model and selecting the model with the lowest minimized VFE.

The difference between VFE and negative log-evidence (NLE) is equal to the Kullback–Leibler divergence (KLD) [3] from the (perfect) Bayesian posterior distribution to the variational distribution for the latent variables in the model. The KLD can be interpreted as the cost of conducting variational rather than Bayesian inference. Perfect (Bayesian) inference would lead to zero inference costs (KLD = 0), and the KLD increases as the variational posterior diverges further from the Bayesian posterior. As a result, model

development in a variational inference context is a balancing act, where we search for models that have both large amounts of evidence for the data and small inference costs (small KLD). In other words, in a variational inference context, the researcher has two knobs to tune models. The first knob alters the model specification, which affects model evidence. The second knob relates to constraining the search space for the variational posterior, which may affect the inference costs.

In this paper, we are concerned with developing algorithms for tuning the second knob. How do we constrain the range of variational posteriors so as to make variational inferences both tractable and accurate (resulting in low KLD)? We present our framework in the context of a (Forney-style) factor graph representation of the model [4,5]. In that context, variational inference can be understood as an automatable and efficient message passing-based inference procedure [6–8].

Traditional constraints include mean-field [6] and Bethe approximations [9,10]. However, more recently it has become clear how alternative local constraints, such as posterior factorization [11], expectation and chance constraints [12,13], and local Laplace approximation [14], may impact both tractability and inference accuracy, and thereby potentially lead to lower VFE. The main contribution of the current work lies in unifying the various ideas on local posterior constraints into a principled method for deriving variational message passing-based inference algorithms. The proposed method derives existing message passing algorithms, but also supports the development of new message passing variants.

Section 2 reviews Forney-style Factor Graphs (FFGs) and variational inference by minimizing the Bethe Free Energy (BFE). This review is continued in Section 3, where we discuss BFE optimization from a Lagrangian optimization viewpoint. In Appendix A, we include an example to illustrate that the Bayes rule can be derived from Lagrangian optimization with data constraints. Our main contribution lies in Section 4, which provides a rigorous treatment of the effects of imposing local constraints on the BFE and the resulting message update rules. We build upon several previous works that describe how manipulation of (local) constraints and variational objectives can be employed to improve variational approximations in the context of message passing. For example, ref. [12] shows how inference algorithms can be unified in terms of hybrid message passing by Lagrangian constraint manipulation. We extend this view by bringing form (Section 4.2) and factorization constraints (Section 4.1) into a constrained optimization framework. In [15], a high-level recipe for generating message passing algorithms from divergence measures is described. We apply their general recipe in the current work, where we adhere to the view on local stationary points for region-based approximations on general graphs [16]. In Appendix B, we also show that locally stationary solutions are also the global stationary solutions. In Section 5, we develop an algorithm for VFE evaluation in an FFG. In previous work, ref. [17] describes a factor softening approach to evaluate the VFE for models with deterministic factors. We extend this work in Section 5, and show how to avoid factor softening for both free energy evaluation and inference of posteriors. We show an example of how to compute VFE for a deterministic node in Appendix C. A more detailed comparison to related work is given in Section 7.

In the literature, proofs and descriptions of message passing-based inference algorithms are scattered across multiple papers and varying graphical representations, including Bayesian networks [6,18], Markov random fields [16], bi-partite (Tanner) factor graphs [12,17,19] and Forney-style factor graphs (FFGs) [5,11]. In Appendix D, we provide first-principle proofs for a large collection of familiar message passing algorithms in the context of Forney-style factor graphs, which is the preferred framework in the information and communication theory communities [4,20].

## 2. Factor Graphs and the Bethe Free Energy

### 2.1. Terminated Forney-Style Factor Graphs

A Forney-style factor graph (FFG) is an undirected graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with nodes  $\mathcal{V}$  and edges  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ . We denote the neighboring edges of a node  $a \in \mathcal{V}$  by  $\mathcal{E}(a)$ . Vice

versa, for an edge  $i \in \mathcal{E}$ , the notation  $\mathcal{V}(i)$  collects all neighboring nodes. As a notational convention, we index nodes by  $a, b, c$  and edges by  $i, j, k$ , unless stated otherwise. We will mainly use  $a$  and  $i$  as summation indices and use the other indices to refer to a node or edge of interest.

In this paper, we will frequently refer to the notion of a subgraph. We define an edge-induced subgraph by  $\mathcal{G}(i) = (\mathcal{V}(i), i)$ , and a node-induced subgraph by  $\mathcal{G}(a) = (a, \mathcal{E}(a))$ . Furthermore, we denote a local subgraph by  $\mathcal{G}(a, i) = (\mathcal{V}(i), \mathcal{E}(a))$ , which collects all local nodes and edges around  $i$  and  $a$ , respectively.

An FFG can be used to represent a factorized function,

$$f(\mathbf{s}) = \prod_{a \in \mathcal{V}} f_a(\mathbf{s}_a), \tag{1}$$

where  $\mathbf{s}_a$  collects the argument variables of factor  $f_a$ . We assumed that all the factors are positive. In an FFG, a node  $a \in \mathcal{V}$  corresponds to a factor  $f_a$ , and the neighboring edges  $\mathcal{E}(a)$  correspond to the variables  $s_a$  that are the arguments of  $f_a$ .

As an example model, the following factorization (2), the corresponding FFG of which is shown in Figure 1.

$$f(s_1, \dots, s_5) = f_a(s_1) f_b(s_1, s_2, s_3) f_c(s_2) f_d(s_3, s_4, s_5) f_e(s_5). \tag{2}$$

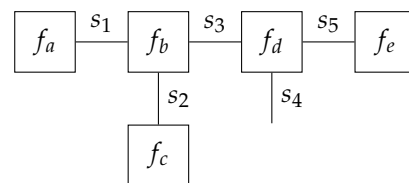


Figure 1. Example Forney-style factor graph for the model of (2).

The FFG of Figure 1 consists of five nodes  $\mathcal{V} = \{a, \dots, e\}$ , as annotated by their corresponding factor functions, and five edges  $\mathcal{E} = \{(a, b), \dots, (d, e)\}$  as annotated by their corresponding variables. An edge that connects to only one node (e.g., the edge for  $s_4$ ) is called a half-edge. In this example, the neighborhood  $\mathcal{E}(b) = \{(a, b), (b, c), (b, d)\}$  and  $\mathcal{V}((b, c)) = \{b, c\}$ .

In the FFG representation, a node can be connected to an arbitrary number of edges, while an edge can only be connected to at most two nodes. Therefore, FFGs often contain “equality nodes” that constrain connected edges to carry identical beliefs, with the implication that these beliefs can be made available to more than two factors. An equality node has the factor function

$$f_a(s_i, s_j, s_k) = \delta(s_j - s_i) \delta(s_j - s_k), \tag{3}$$

for which the node-induced subgraph  $\mathcal{G}(a)$  is drawn in Figure 2.

If every edge in the FFG has exactly two connected nodes (including equality nodes), then we designate the graph as a terminated FFG (TFFG). Since multiplication of a function  $f(\mathbf{s})$  by 1 does not alter the function, any FFG can be terminated by connecting any half-edge  $i$  to a node  $a$  that represents the unity factor  $f_a(s_i) = 1$ .

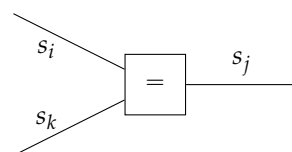


Figure 2. Visualization of the node-induced subgraph for an equality node. If the node function  $f_a$  is known, a symbol representing the node function is often substituted within the node (“=” in this case).

In Section 4.2 we discuss form constraints on posterior distributions. If such a constraint takes on a Dirac-delta functional form, then we visualize the constraint on the FFG by a small circle in the middle of the edge. For example, the small shaded circle in Figure 11 indicates that the variable has been observed. In Section 4.3.2 we consider form constraints in the context of optimization, in which case the circle annotation will be left open (see, e.g., Figure 14).

## 2.2. Variational Free Energy

Given a model  $f(s)$  and a (normalized) probability distribution  $q(s)$ , we can define a Variational Free Energy (VFE) functional as

$$F[q, f] \triangleq \int q(s) \log \frac{q(s)}{f(s)} ds. \quad (4)$$

Variational inference is concerned with finding solutions to the minimization problem

$$q^*(s) = \arg \min_{q \in \mathcal{Q}} F[q, f], \quad (5)$$

where  $\mathcal{Q}$  imposes some constraints on  $q$ .

If  $q$  is unconstrained, then the optimal solution is obtained for  $q^*(s) = p(s)$ , with  $p(s) = \frac{1}{Z} f(s)$  being the exact posterior, and  $Z = \int f(s) ds$  a normalizing constant that is commonly referred to as the evidence. The minimum value of the free energy then follows as the negative log-evidence (NLE),

$$F[q^*, f] = -\log Z,$$

which is also known as the surprisal. The NLE can be interpreted as a measure of model performance, where low NLE is preferred.

As an unconstrained search space for  $q$  grows exponentially with the number of variables, the optimization of (5) quickly becomes intractable beyond the most basic models. Therefore, constraints and approximations to the variational free energy (4) are often utilized. As a result, the *constrained* variational free energy with  $q^* \in \mathcal{Q}$  bounds the NLE by

$$F[q^*, f] = -\log Z + \int q^*(s) \log \frac{q^*(s)}{p(s)} ds, \quad (6)$$

where the latter term expresses the divergence from the (intractable) exact solution to the optimal variational belief.

In practice, the functional form of  $q(s) = q(s; \theta)$  is often parameterized, such that gradients of  $F$  can be derived w.r.t. the parameters  $\theta$ . This effectively converts the variational optimization of  $F[q, f]$  to a parametric optimization of  $F(\theta)$  as a function of  $\theta$ . This problem can then be solved by a (stochastic) gradient descent procedure [21,22].

In the context of variational calculus, while form constraints may lead to interesting properties (see Section 4.2), they are generally not required. Interestingly, in a variational optimization context, the functional form of  $q$  is often not an *assumption*, but rather a *result* of optimization (see Section 4.3.1). An example of variational inference is provided in Appendix A.

## 2.3. Bethe Free Energy

The Bethe approximation enjoys a unique place in the landscape of  $\mathcal{Q}$ , because the Bethe free energy (BFE) defines the fundamental objective of the celebrated belief propagation (BP) algorithm [17,23]. The origin of the Bethe approximation is rooted in tree-like approximations to subgraphs (possibly containing cycles) by enforcing local consistency conditions on the beliefs associated with edges and nodes [24].

Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  for a factorized function  $f(\mathbf{s}) = \prod_{a \in \mathcal{V}} f_a(\mathbf{s}_a)$  (1), the Bethe free energy (BFE) is defined as [25]:

$$F[q, f] \triangleq \sum_{a \in \mathcal{V}} \underbrace{\int q_a(\mathbf{s}_a) \log \frac{q_a(\mathbf{s}_a)}{f_a(\mathbf{s}_a)} d\mathbf{s}_a}_{F[q_a, f_a]} + \sum_{i \in \mathcal{E}} \underbrace{\int q_i(s_i) \log \frac{1}{q_i(s_i)} ds_i}_{H[q_i]} \quad (7)$$

such that the factorized beliefs

$$q(\mathbf{s}) = \prod_{a \in \mathcal{V}} q_a(\mathbf{s}_a) \prod_{i \in \mathcal{E}} q_i(s_i)^{-1} \quad (8)$$

satisfy the following constraints:

$$\int q_a(\mathbf{s}_a) d\mathbf{s}_a = 1, \quad \text{for all } a \in \mathcal{V} \quad (9a)$$

$$\int q_a(\mathbf{s}_a) d\mathbf{s}_{a \setminus i} = q_i(s_i), \quad \text{for all } a \in \mathcal{V} \text{ and all } i \in \mathcal{E}(a). \quad (9b)$$

Together, the normalization constraint (9a) and marginalization constraint (9b) imply that the edge marginals are also normalized:

$$\int q_i(s_i) ds_i = 1, \quad \text{for all } i \in \mathcal{E}. \quad (10)$$

The Bethe free energy (7) includes a local free energy term  $F[q_a, f_a]$  for each node  $a \in \mathcal{V}$ , and an entropy term  $H[q_i]$  for each edge  $i \in \mathcal{E}$ . Note that the local free energy also depends on the node function  $f_a$ , as specified in the factorization of  $f$  (1), whereas the entropy only depends on the local belief  $q_i$ .

The Bethe factorization (8) and constraints are summarized by the local polytope [26]

$$\mathcal{L}(\mathcal{G}) = \{q_a \text{ for all } a \in \mathcal{V} \text{ s.t. (9a), and } q_i \text{ for all } i \in \mathcal{E}(a) \text{ s.t. (9b)}\}, \quad (11)$$

which defines the constrained search space for the factorized variational distribution (8).

#### 2.4. Problem Statement

In this paper, the problem is to find the beliefs in the local polytope that minimize the Bethe free energy

$$q^*(\mathbf{s}) = \arg \min_{q \in \mathcal{L}(\mathcal{G})} F[q, f], \quad (12)$$

where  $q$  is defined by (8), and where  $q \in \mathcal{L}(\mathcal{G})$  offers a shorthand notation for optimizing over the individual beliefs in the local polytope. In the following sections, we will follow the Lagrangian optimization approach to derive various message passing-based inference algorithms.

#### 2.5. Sketch of Solution Approach

The problem statement of Section 2.4 defines a global minimization of the beliefs in the Bethe factorization. Instead of solving the global optimization problem directly, we employ the factorization of the variational posterior and local polytope to subdivide the global problem statement in multiple *interdependent* local objectives.

From the BFE objective (12) and local polytope of (11), we can construct the Lagrangian

$$\begin{aligned}
L[q, f] = & \sum_{a \in \mathcal{V}} F[q_a, f_a] + \sum_{a \in \mathcal{V}} \psi_a \left[ \int q_a(\mathbf{s}_a) \, \mathbf{d}\mathbf{s}_a - 1 \right] + \sum_{a \in \mathcal{V}} \sum_{i \in \mathcal{E}(a)} \int \lambda_{ia}(s_i) \left[ q_i(s_i) - \int q_a(\mathbf{s}_a) \, \mathbf{d}\mathbf{s}_{a \setminus i} \right] \, \mathbf{d}s_i \\
& + \sum_{i \in \mathcal{E}} H[q_i] + \sum_{i \in \mathcal{E}} \psi_i \left[ \int q_i(s_i) \, \mathbf{d}s_i - 1 \right], \tag{13}
\end{aligned}$$

where the Lagrange multipliers  $\psi_a$ ,  $\psi_i$  and  $\lambda_{ia}$  enforce the normalization and marginalization constraints of (9). It can be seen that this Lagrangian contains local beliefs  $q_a$  and  $q_i$ , which are coupled through the  $\lambda_{ia}$  Lagrange multipliers. The Lagrange multipliers  $\lambda_{ia}$  are doubly indexed, because there is a multiplier associated with each marginalization constraint. The Lagrangian method then converts a constrained optimization problem of  $F[q, f]$  to an unconstrained optimization problem of  $L[q, f]$ . The total variation of the Lagrangian (13) can then be approached from the perspective of variations of the individual (coupled) local beliefs.

More specifically, given a locally connected pair  $b \in \mathcal{V}, j \in \mathcal{E}(b)$ , we can rewrite the optimization of (12) in terms of the local beliefs  $q_b, q_j$ , and the constraints in the local polytope

$$\mathcal{L}(\mathcal{G}(b, j)) = \{q_b \text{ s.t. (9a), and } q_j \text{ s.t. (9b)}\}, \tag{14}$$

that pertains to these beliefs. The problem then becomes finding local stationary solutions

$$\{q_b^*, q_j^*\} = \arg \min_{\mathcal{L}(\mathcal{G}(b, j))} F[q, f]. \tag{15}$$

Using (13), the optimization of (15) can then be written in the Lagrangian form

$$q_b^* = \arg \min_{q_b} L_b[q_b, f_b], \tag{16a}$$

$$q_j^* = \arg \min_{q_j} L_j[q_j], \tag{16b}$$

where the Lagrangians  $L_b$  and  $L_j$  include the local polytope of (14) to rewrite (13) as an explicit functional of beliefs  $q_b$  and  $q_j$  (see, e.g., Lemmas 1 and 2). The combined stationary solutions to the local objectives then also comprise a stationary solution to the global objective (Appendix B).

The current paper shows how to identify stationary solutions to local objectives of the form (15), with the use of variational calculus, under varying constraints as imposed by the local polytope (14). Interestingly, the resulting fixed-point equations can be interpreted as message passing updates on the underlying TFFG representation of the model. In the following Sections 3 and 4, we derive the local stationary solutions under a selection of constraints and show how these relate to known message passing update rules (Table 1). It then becomes possible to derive novel message updates and algorithms by simply altering the local polytope.

**Table 1.** Relation between local constraints and derived message updates. The rows refer to different constraints that relate to factor–variable combinations, factors, and variables, respectively. Note that each message passing algorithm combines a set of constraints. Abbreviations: Sum-Product (SP), Structured Variational Message Passing (SVMP), Mean-Field Variational Message Passing (MFVMP), Data Constraint (DC), Laplace Propagation (LP), Mean-Field Variational Laplace (MFVLP), Expectation Maximization (EM), and Expectation Propagation (EP).

Local Constraint	SP	SVMP	MFVMP	DC	LP	MFVLP	EM	EP
Normalization	✓	✓	✓	✓	✓	✓	✓	✓
Marginalization	✓	✓	✓	✓	✓	✓	✓	✓
Moment-Matching								✓
Structured Mean-Field		✓					✓	
Naive Mean-Field			✓			✓		
Laplace Approximation					✓	✓		
Dirac-delta Estimation				✓			✓	✓

### 3. Bethe Lagrangian Optimization by Message Passing

#### 3.1. Stationary Points of the Bethe Lagrangian

We wish to minimize the Bethe free energy under variations of the variational density. As the Bethe free energy factorizes over factors and variables (7), we first consider variations on separate node- and edge-induced subgraphs.

**Lemma 1.** Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the node-induced subgraph  $\mathcal{G}(b)$  (Figure 3). The stationary points of the Lagrangian (16a) as a functional of  $q_b$ ,

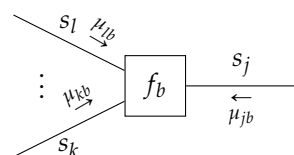
$$L_b[q_b, f_b] = F[q_b, f_b] + \psi_b \left[ \int q_b(\mathbf{s}_b) d\mathbf{s}_b - 1 \right] + \sum_{i \in \mathcal{E}(b)} \int \lambda_{ib}(s_i) \left[ q_i(s_i) - \int q_b(\mathbf{s}_b) ds_{b \setminus i} \right] ds_i + C_b, \tag{17}$$

where  $C_b$  collects all terms that are independent of  $q_b$ , which are of the form

$$q_b(\mathbf{s}_b) = \frac{f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i)}{\int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i) ds_b}. \tag{18}$$

**Proof.** See Appendix D.1. □

The  $\mu_{ib}(s_i)$  are any set of positive functions that makes (18) satisfy (9b), and will be identified in Theorem 1.



**Figure 3.** The subgraph around node  $b$  with indicated messages. Ellipses indicate an arbitrary (possibly zero) amount of edges.

**Lemma 2.** Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider an edge-induced subgraph  $\mathcal{G}(j)$  (Figure 4). The stationary points of the Lagrangian (16b) as a functional of  $q_j$ ,

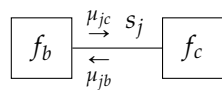


$$L_j[q_j] = H[q_j] + \psi_j \left[ \int q_j(s_j) ds_j - 1 \right] + \sum_{a \in \mathcal{V}(j)} \int \lambda_{ja}(s_j) \left[ q_j(s_j) - \int q_a(s_a) ds_{a \setminus j} \right] ds_j + C_j, \tag{19}$$

where  $C_j$  collects all terms that are independent of  $q_j$ , are of the form

$$q_j(s_j) = \frac{\mu_{jb}(s_j)\mu_{jc}(s_j)}{\int \mu_{jb}(s_j)\mu_{jc}(s_j) ds_j}. \tag{20}$$

**Proof.** See Appendix D.2.  $\square$



**Figure 4.** An edge-induced subgraph  $\mathcal{G}(j)$  with indicated messages.

### 3.2. Minimizing the Bethe Free Energy by Belief Propagation

We now combine Lemmas 1 and 2 to derive the sum-product message update.

**Theorem 1 (Sum-Product Message Update).** Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  (Figure 5). Given the local polytope  $\mathcal{L}(\mathcal{G}(b, j))$  of (14), then the local stationary solutions to (15) are given by

$$q_b^*(s_b) = \frac{f_b(s_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(s_i)}{\int f_b(s_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(s_i) ds_b} \tag{21a}$$

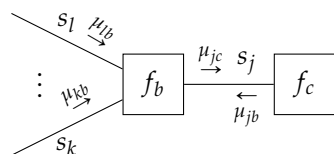
$$q_j^*(s_j) = \frac{\mu_{jb}^*(s_j)\mu_{jc}^*(s_j)}{\int \mu_{jb}^*(s_j)\mu_{jc}^*(s_j) ds_j}, \tag{21b}$$

with messages  $\mu_{jc}^*(s_j)$  corresponding to the fixed points of

$$\mu_{jc}^{(k+1)}(s_j) = \int f_b(s_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}^{(k)}(s_i) ds_{b \setminus j}, \tag{22}$$

with  $k$  representing an iteration index.

**Proof.** See Appendix D.3.  $\square$



**Figure 5.** Visualization of a subgraph with indicated sum-product messages.

The sum-product algorithm has proven to be useful in many engineering applications and disciplines. For example, it is widely used for decoding in communication systems [4,20,27]. Furthermore, for a linear Gaussian state space model, Kalman filtering



and smoothing can be expressed in terms of sum-product message passing for state inference on a factor graph [28,29]. This equivalence has inspired applications ranging from localization [30] to estimation [31].

The sum-product algorithm with updates (22) obtains the exact Bayesian posterior when the underlying graph is a tree [24,25,32]. Application of the sum-product algorithm to cyclic graphs is not guaranteed to converge and might lead to oscillations in the BFE over iterations. Theorems 3.1 and 3.2 in [33] show that the BFE of a graph with a single cycle is convex, which implies that the sum-product algorithm will converge in this case. Moreover, ref. [19] shows that it is possible to obtain a double-loop message passing algorithm if the graph has a cycle such that the stable fixed points will correspond to local minima of the BFE.

**Example 1.** A Linear Dynamical System Considering a Linear Gaussian state space model specified by the following factors:

$$g_0(x_0) = \mathcal{N}(x_0|m_{x_0}, V_{x_0}) \tag{23a}$$

$$g_t(x_{t-1}, z_t, A_t) = \delta(z_t - A_t x_{t-1}) \tag{23b}$$

$$h_t(x'_t, z_t, Q_t) = \mathcal{N}(x'_t|z_t, Q_t^{-1}) \tag{23c}$$

$$n_t(x_t, x'_t, x''_t) = \delta(x_t - x'_t)\delta(x_t - x''_t) \tag{23d}$$

$$m_t(o_t, x''_t, B_t) = \delta(o_t - B_t x''_t) \tag{23e}$$

$$r_t(y_t, o_t, R_t) = \mathcal{N}(y_t|o_t, R_t^{-1}). \tag{23f}$$

The FFG corresponding to the one time segment of the state space model is given in Figure 6. We assumed that we know the following matrices that are used to generate the data:

$$\hat{A}_t = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}, \hat{Q}_t^{-1} = \begin{bmatrix} 3 & 0.1 \\ 0.1 & 2 \end{bmatrix}, \hat{B}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \hat{R}_t^{-1} = \begin{bmatrix} 10 & 2 \\ 2 & 20 \end{bmatrix} \tag{24}$$

with  $\theta = \pi/8$ . Given a collection of observations  $\hat{\mathbf{y}} = \{\hat{y}_1, \dots, \hat{y}_T\}$ , we constrain the latent states  $\mathbf{x} = \{x_0, \dots, x_T\}$  by local marginalization and normalization constraints (for brevity we omit writing the normalization constraints explicitly) in accordance with Theorem 1, i.e.,

$$\int q(x_{t-1}, z_t, A_t) dx_{t-1} dz_t = q(A_t), \int q(x_{t-1}, z_t, A_t) dA_t = q(z_t|x_{t-1})q(x_{t-1}) \tag{25a}$$

$$\int q(x'_t, z_t, Q_t) dx'_t dz_t = q(Q_t), \int q(x'_t, z_t, Q_t) dz_t dQ_t = q(x'_t), \int q(x'_t, z_t, Q_t) dx'_t dQ_t = q(z_t) \tag{25b}$$

$$q(x_t, x'_t, x''_t) = q(x_t)\delta(x_t - x'_t)\delta(x_t - x''_t) \tag{25c}$$

$$\int q(o_t, x''_t, B_t) do_t dx''_t = q(B_t), \int q(o_t, x''_t, B_t) dB_t = q(o_t|x''_t)q(x''_t) \tag{25d}$$

$$\int q(o_t, y_t, R_t) do_t dy_t = q(R_t), \int q(o_t, y_t, R_t) dR_t do_t = q(y_t), \int q(o_t, y_t, R_t) dR_t dy_t = q(o_t) \tag{25e}$$

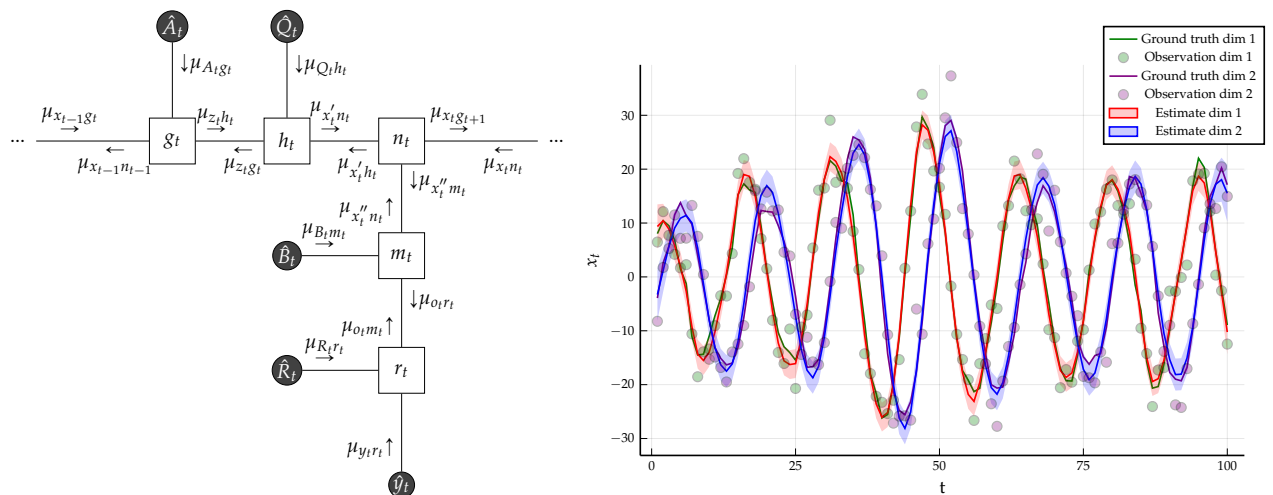
Moreover, we use data constraints in accordance with Theorem 3 (explained in Section 4.2.1) for the observations, state transition matrices and precision matrices, i.e.,

$$q(y_t) = \delta(y_t - \hat{y}_t), q(A_t) = \delta(A_t - \hat{A}_t), q(B_t) = \delta(B_t - \hat{B}_t), q(Q_t) = \delta(Q_t - \hat{Q}_t), q(R_t) = \delta(R_t - \hat{R}_t).$$

Computation of sum-product messages by (22) is analytically tractable and detailed algebraic manipulation can be found in [31]. If the backwards messages are not passed, then the resulting sum-product message passing algorithm is equivalent to Kalman filtering and if both forward and backward messages are propagated, then the Rauch–Tung–Striebel smoother is obtained [34] (Ch. 8).

We generated  $T = 100$  observations  $\hat{\mathbf{y}}$  using the matrices specified in (24) and the initial condition  $\hat{x}_0 = [5, -5]^T$ . Due to (23a), we have  $\mu_{x_0|g_1} = \mathcal{N}(m_{x_0}, V_{x_0})$ . We chose  $V_{x_0} = 100 \cdot I$  and  $m_{x_0} = \hat{x}_0$ . Under these constraints, the results of sum-product message passing and Bethe free

energy evaluation is given in Figure 6. As the underlying graph is a tree, sum-product message passing results are exact and the evaluated BFE corresponds to negative log-evidence. In the follow-up Example 2, we will modify the constraints and give a comparative free energy plot for the examples in Figures 10 and 16.



**Figure 6.** (Left) One time segment of the FFG corresponding to the linear Gaussian state space model specified in Example 1, with the sum-product messages computed according to (22). The three small dots at both sides of the graph indicate identical continuation of the graph over time. (Right) The small dots indicate the noisy observations that are synthetically generated by the linear state space model of (23) using parameter matrices as specified in (24). The posterior distribution for the hidden states are inferred by sum-product message passing and are drawn with shaded regions, indicating plus and minus the variance. The Bethe free energy evaluates to  $F[q, f] = 580.698$ .

#### 4. Message Passing Variations through Constraint Manipulation

For generic node functions with arbitrary connectivity, there is no guarantee that the sum-product updates can be solved analytically. When analytic solutions are not possible, there are two ways to proceed. One way is to try to solve the sum-product update equations numerically, e.g., by Monte Carlo methods. Alternatively, we can add additional constraints to the BFE that leads to simpler update equations at the cost of inference accuracy. In the remainder of the paper, we explore a variety of constraints that have proven to yield useful inference solutions.

##### 4.1. Factorization Constraints

Additional factorizations of the variational density  $q_a(s_a)$  are often assumed to ease computation. In particular, we assumed a *structured mean-field factorization* such that

$$q_b(s_b) \triangleq \prod_{n \in l(b)} q_b^n(s_b^n), \tag{26}$$

where  $n$  indicates a local cluster as a set of edges. To define a local cluster rigorously, let us first denote by  $\mathcal{P}(a)$  the power set of an edge set  $\mathcal{E}(a)$ , where the power set is the set of all subsets of  $\mathcal{E}(a)$ . Then, a mean-field factorization  $l(a) \subseteq \mathcal{P}(a)$  can be chosen such that all elements in  $\mathcal{E}(a)$  are included in  $l(a)$  exactly once. Therefore,  $l(a)$  is defined as a set of one or multiple sets of edges. For example, if  $\mathcal{E}(a) = \{i, j, k\}$ , then  $l(a) = \{\{i\}, \{j, k\}\}$  is allowed, as is  $l(a) = \{\{i, j, k\}\}$  itself, but  $l(a) = \{\{i, j\}, \{j, k\}\}$  is not allowed, since the element  $j$  occurs twice. More formally, in (26), the intersection of the super- and subscript collects the required variables, see Figure 7 for an example. The special case of a fully factorized  $l(b)$  for all edges  $i \in \mathcal{E}(b)$  is known as the *naive mean-field factorization* [11,24].

We will analyze the effect of a structured mean-field factorization (26) on the Bethe free energy (7) for a specific factor node  $b \in \mathcal{V}$ . Substituting (26) in the local free energy for factor  $b$  yields

$$F[q_b, f_b] = F[\{q_b^n\}, f_b] = \sum_{n \in l(b)} \int q_b^n(\mathbf{s}_b^n) \log q_b^n(\mathbf{s}_b^n) d\mathbf{s}_b^n - \int \left\{ \prod_{n \in l(b)} q_b^n(\mathbf{s}_b^n) \right\} \log f_b(\mathbf{s}_b) d\mathbf{s}_b. \quad (27)$$

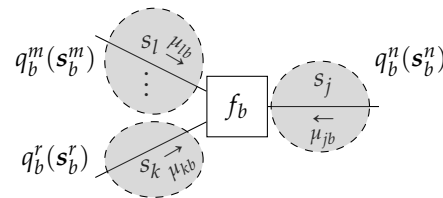
We are then interested in

$$q_b^{m,*} = \arg \min_{q_b^m} L_b^m[q_b^m, f_b], \quad (28)$$

where the Lagrangian  $L_b^m$  (Lemma 3) enforces the normalization and marginalization constraints

$$\int q_b^m(\mathbf{s}_b^m) d\mathbf{s}_b^m = 1, \quad (29a)$$

$$\int q_b^m(\mathbf{s}_b^m) d\mathbf{s}_{b \setminus i}^m = q_i(s_i), \text{ for all } i \in m, m \in l(b). \quad (29b)$$



**Figure 7.** A node-induced subgraph  $\mathcal{G}(b)$  with shaded sections that enclose the edges of an exemplary structured mean-field factorization  $l(b) = \{m, n, r\}$ . Note that, in this example, the cluster  $n$  only encompasses the single edge  $j$ , such that  $q_b^n(\mathbf{s}_b^n) = q_j(s_j)$ . In general, the assignment and number of edges in a cluster can be arbitrary.

**Lemma 3.** Given a terminated FFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider a node-induced subgraph  $\mathcal{G}(b)$  with a structured mean-field factorization  $l(b)$  (e.g., Figure 7). Then, local stationary solutions to the Lagrangian

$$L_b^m[q_b^m] = \int q_b^m(\mathbf{s}_b^m) \log q_b^m(\mathbf{s}_b^m) d\mathbf{s}_b^m - \int \left\{ \prod_{n \in l(b)} q_b^n(\mathbf{s}_b^n) \right\} \log f_b(\mathbf{s}_b) d\mathbf{s}_b + \psi_b^m \left[ \int q_b^m(\mathbf{s}_b^m) d\mathbf{s}_b^m - 1 \right] + \sum_{i \in m} \int \lambda_{ib}(s_i) \left[ q_i(s_i) - \int q_b^m(\mathbf{s}_b^m) d\mathbf{s}_{m \setminus i} \right] d\mathbf{s}_i + C_b^m, \quad (30)$$

where  $C_b^m$  collects all terms independent of  $q_b^m$ , which are of the form

$$q_b^m(\mathbf{s}_b^m) = \frac{\tilde{f}_b^m(\mathbf{s}_b^m) \prod_{i \in m} \mu_{ib}(s_i)}{\int \tilde{f}_b^m(\mathbf{s}_b^m) \prod_{i \in m} \mu_{ib}(s_i) d\mathbf{s}_b^m}, \quad (31)$$

where

$$\tilde{f}_b^m(\mathbf{s}_b^m) = \exp \left( \int \left\{ \prod_{\substack{n \in l(b) \\ n \neq m}} q_b^n(\mathbf{s}_b^n) \right\} \log f_b(\mathbf{s}_b) d\mathbf{s}_b \right)^m. \quad (32)$$

**Proof.** See Appendix D.4.  $\square$

### 4.1.1. Structured Variational Message Passing

We now combine Lemmas 2 and 3 to derive the structured variational message passing algorithm.

**Theorem 2.** *Structured variational message passing: Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  with a structured mean-field factorization  $l(b) \subseteq \mathcal{P}(b)$ , with local clusters  $n \in l(b)$ . Let  $m \in l(b)$  be the cluster where  $j \in m$  (see, e.g., Figure 8). Given the local polytope*

$$\mathcal{L}(\mathcal{G}(b, j)) = \{q_b^n \text{ for all } n \in l(b) \text{ s.t. (29a), and } q_j \text{ s.t. (29b)}\}, \tag{33}$$

then local stationary solutions to

$$\{q_b^{m,*}, q_j^*\} = \arg \min_{\mathcal{L}(\mathcal{G}(b, j))} F[q, f], \tag{34}$$

are given by

$$q_b^{m,*}(\mathbf{s}_b^m) = \frac{\tilde{f}_b^{m,*}(\mathbf{s}_b^m) \prod_{i \in m} \mu_{ib}^*(s_i)}{\int \tilde{f}_b^{m,*}(\mathbf{s}_b^m) \prod_{i \in m} \mu_{ib}^*(s_i) d\mathbf{s}_b^m} \tag{35a}$$

$$q_j^*(s_j) = \frac{\mu_{jb}^*(s_j) \mu_{jc}^*(s_j)}{\int \mu_{jb}^*(s_j) \mu_{jc}^*(s_j) ds_j}, \tag{35b}$$

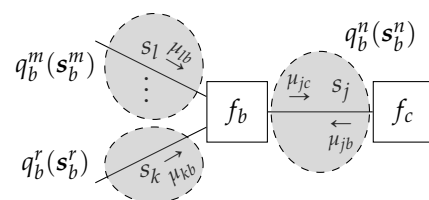
with messages  $\mu_{jc}^*(s_j)$  corresponding to the fixed points of

$$\mu_{jc}^{(k+1)}(s_j) = \int \tilde{f}_b^{m,(k)}(\mathbf{s}_b^m) \prod_{\substack{i \in m \\ i \neq j}} \mu_{ib}^{(k)}(s_i) d\mathbf{s}_{b \setminus j}^m, \tag{36}$$

with iteration index  $k$ , and where

$$\tilde{f}_b^{m,(k)} = \exp \left( \int \left\{ \prod_{\substack{n \in l(b) \\ n \neq m}} q_b^{n,(k)}(\mathbf{s}_b^n) \right\} \log f_b(\mathbf{s}_b) d\mathbf{s}_b^m \right). \tag{37}$$

**Proof.** See Appendix D.5.  $\square$



**Figure 8.** An example subgraph corresponding to  $\mathcal{G}(b, j)$ . Dashed ellipses enclose the edges of an exemplary exact cover  $l(b) = \{m, n, r\}$ . In general, the assignment and number of edges in a cluster can be arbitrary.

The structured mean-field factorization applies the marginalization constraint only to the local cluster beliefs, as opposed to the joint node belief. As a result, computation for the local cluster beliefs might become tractable [24] (Ch.5). The practical appeal of Variational Message Passing (VMP) based inference becomes evident when the underlying model is composed of conjugate factor pairs from the exponential family. When the underlying factors are conjugate exponential family distributions, the message passing updates (36) amounts to adding natural parameters [35] of the underlying exponential family distributions. Structured variational message passing is popular in acoustic signal

modelling, e.g., [36], as it allows one to be able to keep track of correlations over time. In [37], a stochastic variant of structured variational inference is utilized for Latent Dirichlet Allocation. Structured approximations are also used to improve inference in auto-encoders. In [38], inference involving non-parametric Beta-Bernoulli process priors is improved by developing a structured approximation to variational auto-encoders. When the data being modelled are time series, structured approximations reflect the transition structure over time. In [39], an efficient structured black-box variational inference algorithm for fitting Gaussian variational models to latent time series is proposed.

**Example 2.** Consider the linear Gaussian state space model of Example 1. Let us assume that the precision matrix for latent-state transitions  $Q_t$  is not known and can not be constrained by data. Then, we can augment state space model by including a prior for  $Q_t$  and try to infer a posterior over  $Q_t$  from the observations. Since  $Q_t$  is the precision of a normal factor, we chose a conjugate Wishart prior and assumed that  $Q_t$  is time-invariant by adding the following factors

$$w_0(Q_0, V, v) = \mathcal{W}(Q_0|V, v) \tag{38a}$$

$$w_t(Q_{t-1}, Q_t, Q_{t+1}) = \delta(Q_{t-1} - Q_t)\delta(Q_t - Q_{t+1}), \text{ for every } t = 1, \dots, T. \tag{38b}$$

It is certainly possible to assume a time-varying structure for  $Q_t$ ; however, our purpose is to illustrate a change in constraints rather than analyzing time-varying properties. This is why we assume time-invariance.

In this setting, the sum-product equations around the factor  $h_t$  are not analytically tractable. Therefore, we changed the constraints associated with  $h_t$  (25b) to those given in Theorem 2 as follows

$$\int q(x'_t, z_t, Q_t) dx'_t dz_t = q(Q_t), \quad \int q(x'_t, z_t, Q_t) dQ_t = q(x'_t, z_t) \tag{39a}$$

$$\int q(Q_t) dQ_t = 1, \quad \int q(x'_t, z_t) dx'_t dz_t = 1. \tag{39b}$$

We removed the data constraint on  $q(Q_t)$  and instead included data constraints on the hyper-parameters

$$q(V) = \delta(V - \hat{V}), \quad q(v) = \delta(v - \hat{v}). \tag{40}$$

With the new set of constraints ((39a) and (39b)), we obtained a hybrid of the sum-product and structured VMP algorithm, where structured messages around the factor  $h_t$  are computed by (36) and the rest of the messages are computed by the sum-product (22). One time segment of the modified FFG along with the messages is given Figure 9. We used the same observations  $\mathbf{y}$  that were generated in Example 1 and the same initialization for the hidden states. For the hyper-parameters of the Wishart prior, we chose  $\hat{V} = 0.1 \cdot \mathbf{I}$  and  $\hat{v} = 2$ . Under these constraints, the result of structured variational message passing results along with the Bethe free energy evaluation is given in Figure 9.

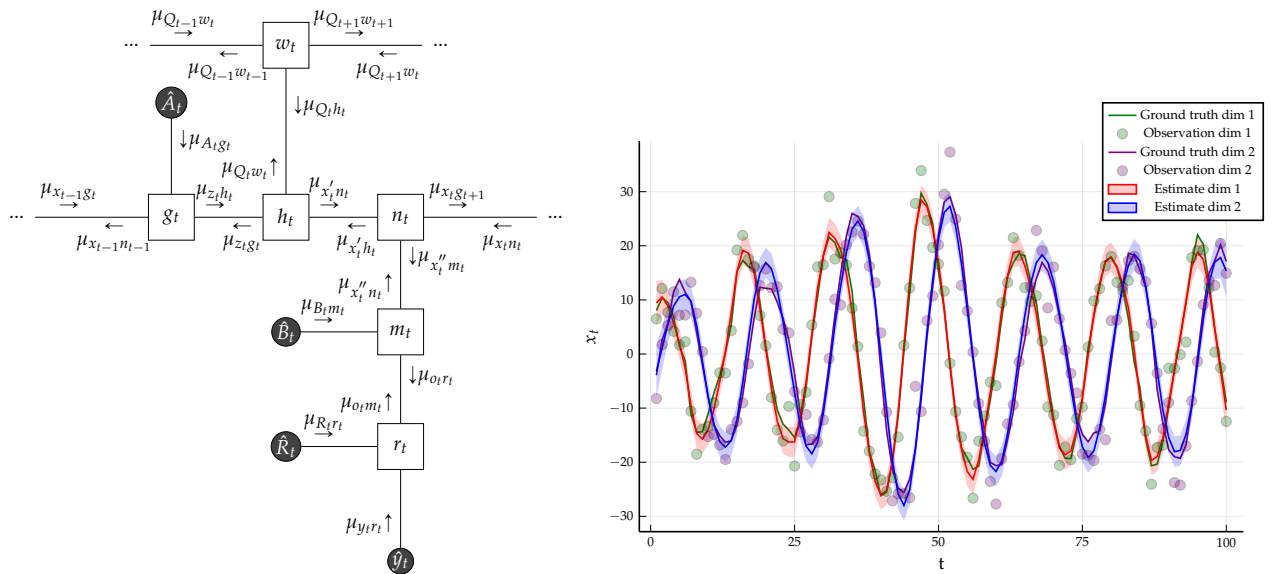
#### 4.1.2. Naive Variational Message Passing

As a corollary of Theorem 2, we can consider the special case of a naive mean-field factorization, which is defined for node  $b$  as

$$q_b(\mathbf{s}_b) = \prod_{i \in \mathcal{E}(b)} q_i(s_i). \tag{41}$$

The naive mean-field constraint (41) transforms the local free energy into

$$\begin{aligned} F[q_b, f_b] &= F[\{q_i\}, f_b] \\ &= \sum_{i \in \mathcal{E}(b)} \int q_i(s_i) \log q_i(s_i) ds_i - \int \left\{ \prod_{i \in \mathcal{E}(b)} q_i(s_i) \right\} \log f_b(\mathbf{s}_b) d\mathbf{s}_b. \end{aligned} \tag{42}$$



**Figure 9.** (Left) One time segment of the FFG corresponding to the linear Gaussian state space model specified in Example 2 with the sum-product messages computed according to (36). (Right) The small dots indicate the noisy observations that are synthetically generated by the linear state space model of (23) using matrices specified in (24). The posterior distribution of the hidden states inferred by structured variational message passing is depicted with shaded regions representing plus and minus one variances. The minimum of the evaluated Bethe free energy over all iterations is  $F[q, f] = 586.178$  (compared to  $F[q, f] = 580.698$  in Example 1). The posterior distribution for the precision matrix is given by  $Q \sim \mathcal{W}\left(\begin{bmatrix} 0.00266 & 0.000334 \\ 0.00034 & 0.00670 \end{bmatrix}, 102.0\right)$ .

**Corollary 1.** Naive Variational Message Passing: Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  with a naive mean-field factorization  $l(b) = \{i \text{ such that for all } i \in \mathcal{E}(b)\}$ . Let  $m \in l(b)$  be the cluster where  $j = m$ . Given the local polytope of (33), the local stationary solutions to (34) are given by

$$q_b^{m,*}(s_b^m) = q_j^*(s_j) = \frac{\mu_{j_b}^*(s_j)\mu_{j_c}^*(s_j)}{\int \mu_{j_b}^*(s_j)\mu_{j_c}^*(s_j) ds_j},$$

where the messages  $\mu_{j_c}^*(s_j)$  are the fixed points of the following iterations

$$\mu_{j_c}^{(k+1)}(s_j) = \exp\left(\int \left\{ \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} q_i^{(k)}(s_i) \right\} \log f_b(s_b) ds_{b \setminus j}\right), \tag{43}$$

where  $k$  is an iteration index.

**Proof.** See Appendix D.6.  $\square$

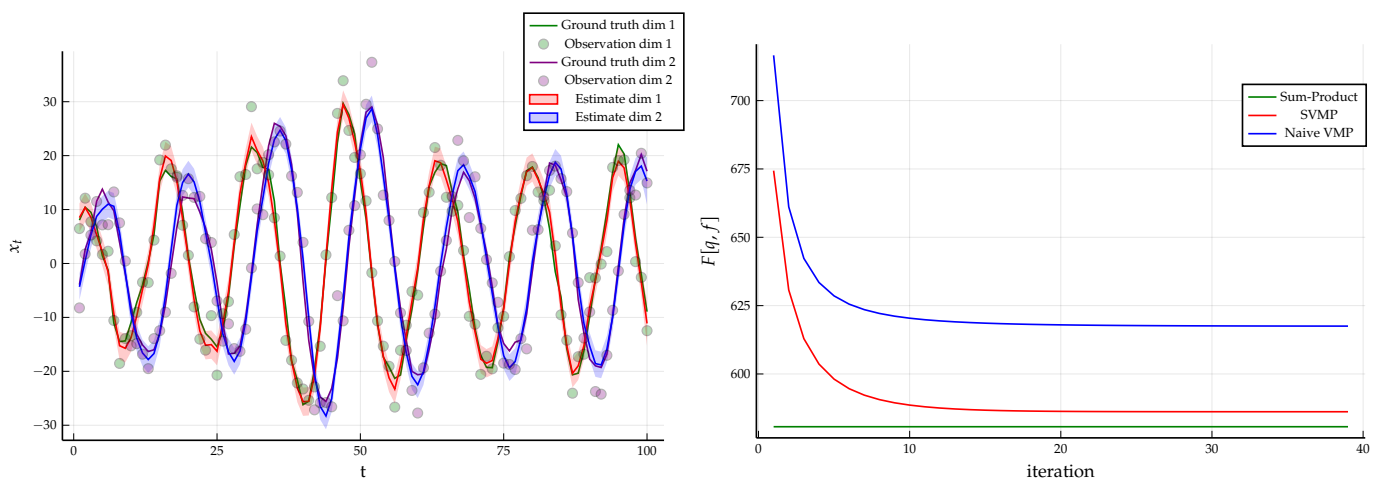
The naive mean-field factorization limits the search space of beliefs by imposing strict constraints on the variational posterior. As a result, the variational posterior also loses flexibility. To improve inference performance for sparse Bayesian learning, the authors of [40] proposes a hybrid mechanism by augmenting naive mean-field VMP with sum-product updates. This hybrid scheme reduces the complexity of the sum-product algorithm, while improving the accuracy of the naive VMP approach. In [41], naive VMP is applied to semi-parametric regression and allows for scaling of regression models to large data sets.

**Example 3.** As a follow up on Example 2, we relaxed the constraints in ((39a) and (39b)) to the following constraints presented in Corollary 1 as

$$\int q(x'_t, z_t, Q_t) dx'_t dz_t = q(Q_t), \int q(x'_t, z_t, Q_t) dQ_t = q(x'_t, z_t) = q(x'_t)q(z_t) \quad (44a)$$

$$\int q(Q_t) dQ_t = 1, \int q(x'_t) dx'_t = 1, \int q(z_t) dz_t = 1. \quad (44b)$$

The FFG remains the same and we use identical data constraints as in Example 2. Together with constraint (44), we obtained a hybrid of naive variational message passing and sum-product message passing algorithm where the messages around the factor  $h_t$  are computed by (43) and the rest of the messages by sum-product (22). Using the same data as in Example 1, the results for naive VMP are given in Figure 10 along with the evaluated Bethe free energy.



**Figure 10.** (Left) The small dots indicate the noisy observations that were synthetically generated by the linear state space model of (23) using matrices specified in (24). The posterior distribution for the hidden states inferred by naive variational message passing is depicted with shaded regions representing plus and minus one variances. The minimum of the evaluated Bethe free energy over all iterations is  $F[q, f] = 617.468$ , which is more than for the less-constrained Example 2 (with  $F[q, f] = 586.178$ ) and Example 1 (with  $F[q, f] = 580.698$ ). The posterior for the precision matrix is given by  $Q \sim \mathcal{W}\left(\begin{bmatrix} 0.00141 & -6.00549e^{-5} \\ -6.00549e^{-5} & 0.00187 \end{bmatrix}, 102.0\right)$ . (Right) A comparison of the Bethe free energies for sum-product, structured and naive variational message passing algorithms for the data generated in Example 1.

#### 4.2. Form Constraints

Form constraints limit the functional form of the variational factors  $q_a(s_a)$  and  $q_i(s_i)$ . One of the most widely used form constraints, the data constraint, is also illustrated in Appendix A.

##### 4.2.1. Data Constraints

A data constraint can be viewed as a special case of (9b), where the belief  $q_j$  is constrained to be a Dirac-delta function [42], such that

$$\int q_a(s_a) ds_{a \setminus j} = q_j(s_j) = \delta(s_j - \hat{s}_j), \quad (45)$$

where  $\hat{s}_j$  is a known value, e.g., an observation.

**Lemma 4.** Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the node-induced subgraph  $\mathcal{G}(b)$  (Figure 3). Then local stationary solutions to the Lagrangian



$$L_b[q_b, f_b] = F[q_b, f_b] + \psi_b \left[ \int q_b(\mathbf{s}_b) d\mathbf{s}_b - 1 \right] + \sum_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \int \lambda_{ib}(s_i) \left[ q_i(s_i) - \int q_b(\mathbf{s}_b) d\mathbf{s}_{b \setminus i} \right] ds_i + \int \lambda_{jb}(s_j) \left[ \delta(s_j - \hat{s}_j) - \int q_b(\mathbf{s}_b) d\mathbf{s}_{b \setminus j} \right] ds_j + C_b. \tag{46}$$

where  $C_b$  collects all terms that are independent of  $q_b$ , are of the form

$$q_b(\mathbf{s}_b) = \frac{f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i)}{\int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i) d\mathbf{s}_b}. \tag{47}$$

**Proof.** See Appendix D.7. □

**Theorem 3.** *Data-Constrained Sum-Product:* Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  (Figure 11). Given the local polytope

$$\mathcal{L}(\mathcal{G}(b, j)) = \{q_b \text{ s.t. (45)}\}, \tag{48}$$

the local stationary solutions to

$$q_b^* = \arg \min_{\mathcal{L}(\mathcal{G}(b, j))} F[q, f],$$

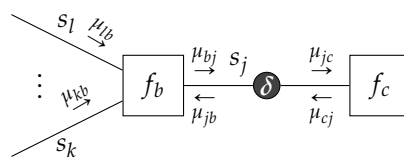
are of the form

$$q_b^*(\mathbf{s}_b) = \frac{f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(s_i)}{\int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(s_i) d\mathbf{s}_b}, \tag{49}$$

with message

$$\mu_{jb}^*(s_j) = \delta(s_j - \hat{s}_j). \tag{50}$$

**Proof.** See Appendix D.8. □



**Figure 11.** Visualization of a subgraph  $\mathcal{G}(b, j)$  with indicated messages, where the dark circled delta indicates a data constraint—i.e., the variable  $s_j$  is constrained to have a distribution of the form  $\delta(s_j - \hat{s}_j)$ .

Note that the resulting message  $\mu_{jb}^*(s_j)$  to node  $b$  does not depend on messages from node  $c$ , as would be the case for a sum-product update. By the symmetry of Theorem 3 for the subgraph  $\mathcal{L}\{\mathcal{G}(c, j)\}$ , (A32) identifies

$$\mu_{cj}(s_j) = \int f_c(\mathbf{s}_c) \prod_{\substack{i \in \mathcal{E}(c) \\ i \neq j}} \mu_{ic}(s_i) d\mathbf{s}_{c \setminus j} \neq \delta(s_j - \hat{s}_j).$$

This implies that messages incoming to a data constraint (such as  $\mu_{cj}$ ) are not further propagated through the data constraint. The data constraint thus effectively introduces a conditional independence between the variables of neighboring factors (conditioned on the shared constrained variable). Interestingly, this is similar to the notion of an intervention [43], where a decision variable is externally forced to a realization.

Data constraints allow information from data sets to be absorbed into the model. Essentially, (variational) Bayesian machine learning is an application of inference in a graph with data constraints. In our framework, data are a constraint, and machine learning via Bayes rule follows naturally from the minimization of the Bethe free energy (see also Appendix A).

#### 4.2.2. Laplace Propagation

A second type of form constraint we consider is the Laplace constraint, see also [14]. Consider a second-order Taylor approximation on the local log-node function

$$\mathcal{L}_a(\mathbf{s}_a) = \log f_a(\mathbf{s}_a), \tag{51}$$

around an approximation point  $\hat{\mathbf{s}}_a$ , as

$$\tilde{\mathcal{L}}_a(\mathbf{s}_a; \hat{\mathbf{s}}_a) = \mathcal{L}_a(\hat{\mathbf{s}}_a) + \nabla^\top \mathcal{L}_a(\hat{\mathbf{s}}_a)(\mathbf{s}_a - \hat{\mathbf{s}}_a) + \frac{1}{2}(\mathbf{s}_a - \hat{\mathbf{s}}_a)^\top \nabla^2 \mathcal{L}_a(\hat{\mathbf{s}}_a)(\mathbf{s}_a - \hat{\mathbf{s}}_a). \tag{52}$$

From this approximation, we define the Laplace-approximated node function as

$$\tilde{f}_a(\mathbf{s}_a; \hat{\mathbf{s}}_a) \triangleq \exp(\tilde{\mathcal{L}}_a(\mathbf{s}_a; \hat{\mathbf{s}}_a)), \tag{53}$$

which is substituted in the local free energy to obtain the Laplace-encoded local free energy as

$$F[q_a, \tilde{f}_a; \hat{\mathbf{s}}_a] = \int q_a(\mathbf{s}_a) \log \frac{q_a(\mathbf{s}_a)}{\tilde{f}_a(\mathbf{s}_a; \hat{\mathbf{s}}_a)} d\mathbf{s}_a. \tag{54}$$

It follows that the Laplace-encoded optimization of the local free energy becomes

$$q_a^* = \arg \min_{q_a} L_a[q_a, \tilde{f}_a; \hat{\mathbf{s}}_a], \tag{55}$$

where the Lagrangian  $L_a$  imposes the marginalization and normalization constraints of (9) on (54).

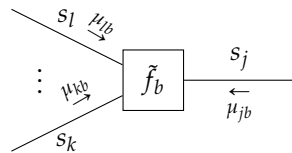
**Lemma 5.** *Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the node-induced subgraph  $\mathcal{G}(b)$  (Figure 12). The stationary points of the Laplace-approximated Lagrangian (55) as a functional of  $q_b$ ,*

$$L_b[q_b, \tilde{f}_b; \hat{\mathbf{s}}_b] = F[q_b, \tilde{f}_b; \hat{\mathbf{s}}_b] + \psi_b \left[ \int q_b(\mathbf{s}_b) d\mathbf{s}_b - 1 \right] + \sum_{i \in \mathcal{E}(b)} \int \lambda_{ib}(s_i) \left[ q_i(s_i) - \int q_b(\mathbf{s}_b) d\mathbf{s}_{b \setminus i} \right] ds_i + C_b, \tag{56}$$

where  $C_b$  collects all terms that are independent of  $q_b$ , which are of the form

$$q_b(\mathbf{s}_b) = \frac{\tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(\mathbf{s}_i)}{\int \tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(\mathbf{s}_i) d\mathbf{s}_b}. \tag{57}$$

**Proof.** See Appendix D.9. □



**Figure 12.** The subgraph around a Laplace-approximated node  $b$  with indicated messages.

We can now formulate Laplace propagation as an iterative procedure, where the approximation point  $\hat{\mathbf{s}}_b$  is chosen as the mode of the belief  $q_b(\mathbf{s}_b)$ .

**Theorem 4.** *Laplace Propagation:* Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  (Figure 13) with the Laplace-encoded factor  $\tilde{f}_b$  as per (53). We write the model (1) with the Laplace-encoded factor  $\tilde{f}_b$  substituted for  $f_b$ , as  $\tilde{f}$ . Given the local polytope  $\mathcal{L}(\mathcal{G}(b, j))$  of (14), the local stationary solutions to

$$\{q_b^*, q_j^*\} = \arg \min_{\mathcal{L}(\mathcal{G}(b, j))} F[q, \tilde{f}; \hat{\mathbf{s}}_b], \tag{58}$$

are given by

$$q_b^*(\mathbf{s}_b) = \frac{\tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b^*) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(\mathbf{s}_i)}{\int \tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b^*) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(\mathbf{s}_i) d\mathbf{s}_b}$$

$$q_j^*(s_j) = \frac{\mu_{jb}^*(s_j) \mu_{jc}^*(s_j)}{\int \mu_{jb}^*(s_j) \mu_{jc}^*(s_j) ds_j},$$

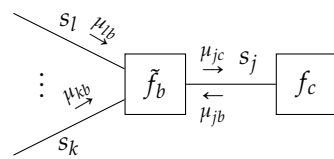
with  $\hat{\mathbf{s}}_b^*$  and the messages  $\mu_{jc}^*(s_j)$  the fixed points of

$$\hat{\mathbf{s}}_b^{(k)} = \arg \max_{\mathbf{s}_b} \log q_b^{(k)}(\mathbf{s}_b)$$

$$q_b^{(k+1)}(\mathbf{s}_b) = \frac{\tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b^{(k)}) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^{(k)}(\mathbf{s}_i)}{\int \tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b^{(k)}) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^{(k)}(\mathbf{s}_i) d\mathbf{s}_b}$$

$$\mu_{jc}^{(k+1)}(s_j) = \int \tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b^{(k)}) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}^{(k)}(\mathbf{s}_i) d\mathbf{s}_{b \setminus j}.$$

**Proof.** See Appendix D.10. □



**Figure 13.** Visualization of a subgraph with indicated Laplace propagation messages. The node function  $f_b$  is denoted by  $\tilde{f}_b$  according to (53).

A Laplace propagation is introduced in [14] as an algorithm that propagates mean and variance information when exact updates are expensive to compute. Laplace propagation has found applications in the context of Gaussian processes and support vector machines [14]. In the jointly normal case, Laplace propagation coincides with sum-product and expectation propagation [14,18].

#### 4.2.3. Expectation Propagation

Expectation propagation can be derived in terms of constraint manipulation by relaxing the marginalization constraints to expectation constraints. Expectation constraints are of the form

$$\int q_a(\mathbf{s}_a) T_i(\mathbf{s}_i) d\mathbf{s}_a = \int q_i(\mathbf{s}_i) T_i(\mathbf{s}_i) d\mathbf{s}_i, \tag{59}$$

for a given function (statistic)  $T_i(\mathbf{s}_i)$ . Technically, the statistic  $T_i(\mathbf{s}_i)$  can be chosen arbitrarily. Nevertheless, they are often chosen as sufficient statistics of an exponential family distribution. An exponential family distribution is defined by

$$q_i(\mathbf{s}_i) = h(\mathbf{s}_i) \exp\left(\eta_i^\top T_i(\mathbf{s}_i) - \log Z(\eta_i)\right), \tag{60}$$

where  $\eta_i$  is the natural parameter,  $Z(\eta_i)$  is the partition function,  $T_i(\mathbf{s}_i)$  is the sufficient statistics and  $h(\mathbf{s}_i)$  is a base measure [24]. The reason  $T_i(\mathbf{s}_i)$  is a sufficient statistic is because if there are observed values of the random variable  $\mathbf{s}_i$ , then the parameter  $\eta_i$  can be estimated by using only the statistics  $T_i(\mathbf{s}_i)$ . This means that the estimator of  $\eta_i$  will depend only on the statistics.

The idea behind expectation propagation [18] is to relax the marginalization constraints with moment-matching constraints by choosing sufficient statistics from exponential family distributions [12]. Relaxation allows approximating the marginals of the sum-product algorithm with exponential family distributions. By keeping the marginals within the exponential family, the complexity of the resulting computations is reduced.

**Lemma 6.** Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the node-induced subgraph  $\mathcal{G}(b)$  (Figure 3). The stationary points of the Lagrangian

$$L_b[q_b, f_b] = F[q_b, f_b] + \psi_b \left[ \int q_b(\mathbf{s}_b) d\mathbf{s}_b - 1 \right] + \sum_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \int \lambda_{ib}(\mathbf{s}_i) \left[ q_i(\mathbf{s}_i) - \int q_b(\mathbf{s}_b) d\mathbf{s}_{b \setminus i} \right] d\mathbf{s}_i + \eta_{jb}^\top \left[ \int q_j(\mathbf{s}_j) T_j(\mathbf{s}_j) d\mathbf{s}_j - \int q_b(\mathbf{s}_b) T_j(\mathbf{s}_j) d\mathbf{s}_b \right] + C_b, \tag{61}$$

with sufficient statistics  $T_j$ , and where  $C_b$  collects all terms that are independent of  $q_b$ , are of the form

$$q_b(\mathbf{s}_b) = \frac{f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(\mathbf{s}_i)}{\int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(\mathbf{s}_i) d\mathbf{s}_b}, \tag{62}$$

with incoming exponential family message

$$\mu_{jb}(s_j) = \exp\left(\eta_{jb}^\top T_j(s_j)\right). \tag{63}$$

**Proof.** See Appendix D.11.  $\square$

**Lemma 7.** Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider an edge-induced subgraph  $\mathcal{G}(j)$  (Figure 4). The stationary solutions of the Lagrangian

$$L_j[q_j] = H[q_j] + \psi_j \left[ \int q_j(s_j) ds_j - 1 \right] + \sum_{a \in \mathcal{V}(j)} \eta_{ja}^\top \left[ \int q_j(s_j) T_j(s_j) ds_j - \int q_a(s_a) T_j(s_j) ds_a \right] + C_j,$$

with sufficient statistics  $T_j(s_j)$ , and where  $C_j$  collects all terms that are independent of  $q_j$ , are of the form

$$q_j(s_j) = \frac{\exp\left([\eta_{jb} + \eta_{jc}]^\top T_j(s_j)\right)}{\int \exp\left([\eta_{jb} + \eta_{jc}]^\top T_j(s_j)\right) ds_j}. \tag{64}$$

**Proof.** See Appendix D.12.  $\square$

**Theorem 5.** Expectation Propagation: Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  (Figure 5). Given the local polytope

$$\mathcal{L}(\mathcal{G}(b, j)) = \{q_b \text{ s.t. (9a), and } q_j \text{ s.t. (59) and (10)}\}, \tag{65}$$

and  $\mu_{jb}(s_j) = \exp\left(\eta_{jb}^\top T_j(s_j)\right)$  an exponential family message (from Lemma 6). Then, the local stationary solutions to (15) are given by

$$q_b^*(s_b) = \frac{f_b(s_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(s_i)}{\int f_b(s_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}^*(s_i) ds_b} \tag{66a}$$

$$q_j^*(s_j) = \frac{\exp\left([\eta_{jb}^* + \eta_{jc}^*]^\top T_j(s_j)\right)}{\int \exp\left([\eta_{jb}^* + \eta_{jc}^*]^\top T_j(s_j)\right) ds_j}, \tag{66b}$$

with  $\eta_{jb}^*, \eta_{jc}^*$  and  $\mu_{jc}^*(s_j)$  being the fixed points of the iterations

$$\tilde{\mu}_{jc}^{(k)}(s_j) = \int f_b(s_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}^{(k)}(s_i) ds_{b \setminus j}$$

$$\tilde{q}_j^{(k)}(s_j) = \frac{\mu_{jb}^{(k)}(s_j) \tilde{\mu}_{jc}^{(k)}(s_j)}{\int \mu_{jb}^{(k)}(s_j) \tilde{\mu}_{jc}^{(k)}(s_j) ds_j}.$$

By moment-matching on  $\tilde{q}_j^{(k)}(s_j)$ , we obtain the natural parameter  $\tilde{\eta}_j^{(k)}$ . The message update then follows from

$$\eta_{jc}^{(k)} = \tilde{\eta}_j^{(k)} - \eta_{jb}^{(k)}$$

$$\mu_{jc}^{(k+1)}(s_j) = \exp\left(T_j(s_j)^\top \eta_{jc}^{(k)}\right).$$

**Proof.** See Appendix D.13.  $\square$

Moment-matching can be performed by solving [24] (Proposition 3.1)

$$\nabla_{\eta_j} \log Z_j(\eta_j) = \int \tilde{q}_j(s_j) T_j(s_j) ds_j$$

for  $\eta_j$ , where

$$Z_j(\eta_j) = \int \exp(\eta_j^\top T_j(s_j)) ds_j.$$

In practice, for a Gaussian approximation, the natural parameters can be obtained by converting the matched mean and variance of  $\tilde{q}_j(s_j)$  to the canonical form [18]. Computing the moments of  $\tilde{q}_j(s_j)$  is often challenging due to lack of closed form solutions of the normalization constant. In order to address the computation of moments in EP, Ref. [44] proposes to evaluate challenging moments by quadrature methods. For multivariate random variables, moment-matching by spherical radial cubature would be advantageous as it will reduce the computational complexity [45]. Another popular way of evaluating the moments is through importance sampling [46] (Ch. 7) and [47].

Expectation propagation has been utilized in various applications ranging from time series estimation with Gaussian processes [48] to Bayesian learning with stochastic natural gradients [49]. When the likelihood functions for Gaussian process classification are not Gaussian, EP is often utilized [50] (Chapter 3). In [51], a message passing-based expectation propagation algorithm is developed for models that involve both continuous and discrete random variables. Perhaps the most practical applications of EP are in the context of probabilistic programming [52], where it is heavily used in real-world applications.

### 4.3. Hybrid Constraints

In this section, we consider hybrid methods that combine factorization and form constraints, and formalize some well-known algorithms in terms of message passing.

#### 4.3.1. Mean-Field Variational Laplace

Mean-field variational Laplace applies the mean-field factorization to the Laplace-approximated factor function. The appeal of this method is that all messages outbound from the Laplace-approximated factor can be represented by Gaussians.

**Theorem 6.** *Mean-field variational Laplace: Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  (Figure 13) with the Laplace-encoded factor  $\tilde{f}_b$  as per (53). We write the model (1) with substituted Laplace-encoded factor  $\tilde{f}_b$  for  $f_b$ , as  $\tilde{f}$ . Furthermore, assume a naive mean-field factorization  $l(b) = \{\{i\} \text{ for all } i \in \mathcal{E}(b)\}$ . Let  $m \in l(b)$  be the cluster where  $j = m$ . Given the local polytope of (33), the local stationary solutions to*

$$\{q_b^{m,*}, q_j^*\} = \arg \min_{\mathcal{L}(\mathcal{G}(b,j))} F[q, \tilde{f}; \hat{s}_b], \tag{67}$$

are given by

$$q_b^{m,*}(s_b^m) = q_j^*(s_j) = \frac{\mu_{jb}^*(s_j) \mu_{jc}^*(s_j)}{\int \mu_{jb}^*(s_j) \mu_{jc}^*(s_j) ds_j},$$

where  $\mu_{jc}^*$  represents the fixed points of the following iterations

$$\mu_{jc}^{(k+1)}(s_j) = \exp \left( \int \left( \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} q_i^{(k)}(s_i) \right) \log \tilde{f}_b(s_b; \hat{s}_b^{(k)}) ds_{b \setminus j} \right), \tag{68}$$

with

$$\hat{s}_b^{(k)} = \arg \max_{s_b} \log q_b^{(k)}(s_b).$$

**Proof.** See Appendix D.14.  $\square$

Conveniently, under these constraints, every outbound message from node  $b$  will be proportional to a Gaussian. Substituting the Laplace-approximated factor function, we obtain:

$$\log \mu_{jc}^{(k)}(s_j) = \int \left( \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} q_i^{(k)}(s_i) \right) \tilde{\mathcal{L}}_b(s_b; \hat{s}_b^{(k)}) ds_{b \setminus j} + C.$$

Resolving this expectation yields a quadratic form in  $s_j$ , which after completing the square leads to a proportionally Gaussian message  $\mu_{jc}(s_j)$ . This argument holds for any edge adjacent to  $b$ , and therefore for all outbound messages from node  $b$ . Moreover, if the incoming messages are represented by Gaussians as well (e.g., because these are also computed under the mean-field variational Laplace constraint), then all beliefs on the adjacent edges to  $b$  will also be Gaussian. This significantly simplifies the procedure of computing the expectations, which illustrates the computational appeal of mean-field variational Laplace.

Mean-field variational Laplace is widely used in dynamic causal modeling [53] and more generally in cognitive neuroscience, partly because the resulting computations are deemed neurologically plausible [54–56].

#### 4.3.2. Expectation Maximization

Expectation Maximization (EM) can be viewed as a hybrid algorithm that combines a structured variational factorization with a Dirac-delta constraint, where the constrained value itself is optimized. Given a structured mean-field factorization  $l(a) \subseteq \mathcal{P}(a)$ , with a single-edge cluster  $m = j$ , then expectation maximization considers local factorizations of the form

$$q_a(s_a) = \delta(s_j - \theta_j) \prod_{\substack{n \in l(a) \\ n \neq m}} q_a^n(s_a^n), \tag{69}$$

where the belief for  $s_j$  is constrained by a Dirac-delta distribution, similar to Section 4.2.1. In (69), however, the variable  $s_j$  represents a random variable with (unknown) value  $\theta_j \in \mathbb{R}^d$ , where  $d$  is the dimension of the random variable  $s_j$ . We explicitly use the notation  $\theta_j$  (as opposed to  $\hat{s}_j$  for the data constraint in Section 4.2.1) to clarify that this value is a parameter for the constrained belief over  $s_j$  that will be optimized—that is,  $\theta_j$  does not represent a model parameter in itself. To make this distinction even more explicit, in the context of optimization, we will refer to Dirac-delta constraints as point-mass constraints.

The factor-local free energy  $F[q_a, f_a; \theta_j]$  then becomes a function of the  $\theta_j$  parameter.

**Theorem 7.** *Expectation maximization: Given a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , consider the induced subgraph  $\mathcal{G}(b, j)$  (Figure 14) with a structured mean-field factorization  $l(b) \subseteq \mathcal{P}(b)$ , with local clusters  $n \in l(b)$ . Let  $m \in l(b)$  be the cluster where  $j = m$ . Given the local polytope*

$$\mathcal{L}(\mathcal{G}(b, j)) = \{q_b^n \text{ for all } n \in l(b) \text{ s.t. (29a)}\}, \tag{70}$$

the local stationary solutions to

$$\theta_j^* = \arg \min_{\mathcal{L}(\mathcal{G}(b, j))} F[q, f; \theta_j],$$

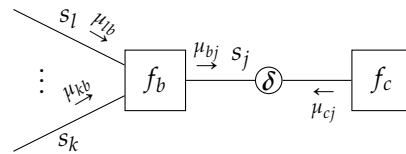


are given by the fixed points of

$$\mu_{bj}^{(k+1)}(s_j) = \exp \left( \int \left\{ \prod_{\substack{n \in I(b) \\ n \neq m}} q_b^{n,(k)}(s_b^n) \right\} \log f_b(s_b) \, ds_{b \setminus j} \right) \tag{71a}$$

$$\theta_j^{(k+1)} = \arg \max_{s_j} \left( \log \mu_{bj}^{(k+1)}(s_j) + \log \mu_{cj}^{(k+1)}(s_j) \right). \tag{71b}$$

**Proof.** See Appendix D.15. □



**Figure 14.** Visualization of a subgraph  $\mathcal{G}(b, j)$  with indicated messages. The open circle indicates a point-mass constraint of the form  $\delta(s_j - \theta_j)$ , where the value  $\theta_j$  is optimized.

Expectation maximization was formulated in [57] as an iterative method that optimizes log-expectations of likelihood functions, where each EM iteration is guaranteed to increase the expected log-likelihood. Moreover, under some differentiability conditions, the EM algorithm is guaranteed to converge [57] (Theorem 3). A detailed overview of EM for exponential families is available in [24] (Ch. 6). A formulation of EM in terms of message passing is given by [58], where message passing for EM is applied in a filtering and system identification context. In [58], derivations are based on [57] (Theorem 1), whereas our derivations directly follow from variational principles.

**Example 4.** Now suppose we do not know the angle  $\theta$  for the state transition matrix  $A_t$  in Example 2 and would like to estimate the value of  $\theta$ . Moreover, further suppose that we are interested in estimating the hyper-parameters for the prior  $m_{x_0}$  and  $V_{x_0}$ , as well as the precision matrix for the state transitions  $Q_t$ . For this purpose, we changed the constraints of (25a) into EM constraints in accordance with Theorem 7:

$$q(x_{t-1}, z_t, A_t(\theta)) = \delta(A_t(\theta) - A_t(\hat{\theta}))q(z_t|x_{t-1}, A_t(\theta))q(x_{t-1}) \tag{72a}$$

$$q(x_0, m_{x_0}, V_{x_0}) = q(x_0)\delta(m_{x_0} - \hat{m}_{x_0})\delta(V_{x_0} - \hat{V}_{x_0}), \tag{72b}$$

where we optimize  $\hat{\theta}$ ,  $\hat{V}_{x_0}$  and  $\hat{m}_{x_0}$  with EM ( $\hat{V}_{x_0}$  is further constrained to be positive definite during the optimization procedure). With the addition of the new EM constraints, the resulting FFG is given in Figure 15. The hybrid message passing algorithm consists of structured variational messages around the factor  $h_t$ , and sum-product messages around  $w_t$ ,  $n_t$ ,  $m_t$  and  $r_t$ , and EM messages around  $g_0$  and  $g_t$ . We used identical observations as in the previous examples. The results for the hybrid SVMP-EM-SP algorithm are given in Figure 16 along with the evaluated Bethe free energy over all iterations.

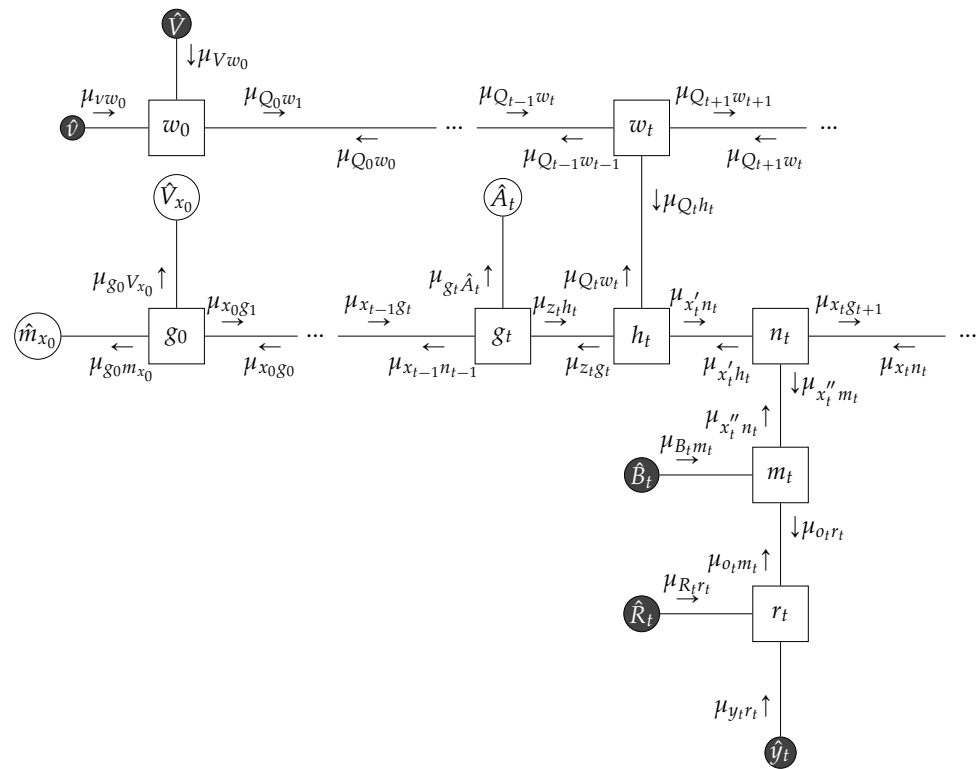


Figure 15. The FFG of the linear Gaussian state space model augmented with the EM constraints in Example 4.

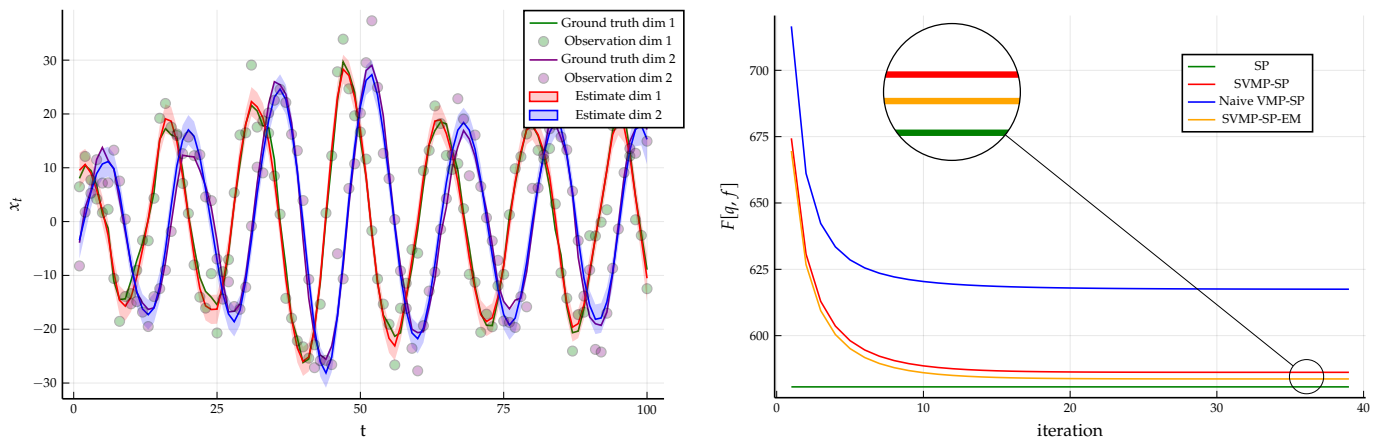


Figure 16. (Left) The small dots indicate the noisy observations that are synthetically generated by the linear state space model of (23) using matrices specified in (24). The posterior distribution of the hidden states inferred by structured variational message passing is depicted with shaded regions representing plus and minus one variances. The minimum of the evaluated Bethe free energy over iterations is  $F[q, f] = 583.683$ . Moreover, the posterior distribution for the precision matrix is given by  $Q \sim \mathcal{W}\left(\begin{bmatrix} 0.00286 & 0.00038 \\ 0.00038 & 0.00691 \end{bmatrix}, 102.0\right)$ . The EM estimates are  $\theta = \pi/7.821$ ,  $\hat{m}_{x_0} = [7.23, -7.016]$  and  $\hat{V}_{x_0} = \begin{bmatrix} 11.028 & -1.926 \\ -1.926 & 10.918 \end{bmatrix}$ . (Right) Free energy plots of the 4 algorithms discussed in Examples 1–4 on the same data set.

4.4. Overview of Message Passing Algorithms

In Sections 4.1–4.3, following a high-level recipe pioneered by [15], we presented first-principle derivations of some of the popular message passing-based inference algorithms by manipulating the local constraints of the Bethe free energy. The results are summarized in Table 1.

Crucially, the method of constrained BFE minimization goes beyond the reviewed algorithms. Through creating a new set of local constraints and following similar derivations based on variational calculus, one can obtain new message passing-based inference algorithms that better match the specifics of the generative model or application.

### 5. Scoring Models by Minimized Variational Free Energy

As discussed in Section 2.2, the variational free energy is an important measure of model performance. In Sections 5.1 and 5.2, we discuss some problems that occur when evaluating the BFE on a TFFG. In Section 5.3, we propose an algorithm that evaluates the constrained BFE as a summation of local contributions on the TFFG.

#### 5.1. Evaluation of the Entropy of Dirac-Delta Constrained Beliefs

For continuous variables, data and point-mass constraints, as discussed in Sections 4.2.1 and 4.3.2 and Appendix A, collapse the information density to infinity, which leads to singularities in entropy evaluation [59]. More specifically, for a continuous variable  $s_j$ , the entropies for beliefs of the form  $q_j(s_j) = \delta(s_j - \hat{s}_j)$  and  $q_a(s_a) = q_{a|j}(s_{a \setminus j} | s_j) \delta(s_j - \hat{s}_j)$  both evaluate to  $-\infty$ .

In variational inference, it is common to define the VFE only with respect to the latent (unobserved) variables [2] (Section 10.1). In contrast, in this paper, we explicitly define the BFE in terms of an iteration over all nodes and edges (7), which also includes non-latent beliefs in the BFE definition. Therefore, we define

$$\begin{aligned} q_j(s_j) &= \delta(s_j - \hat{s}_j) \Rightarrow H[q_j] \triangleq 0, \\ q_a(s_a) &= q_{a|j}(s_{a \setminus j} | s_j) \delta(s_j - \hat{s}_j) \Rightarrow H[q_a] \triangleq H[q_{a \setminus j}], \end{aligned}$$

where  $q_{a|j}(s_{a \setminus j} | s_j)$  indicates the conditional belief and  $q_{a \setminus j}(s_{a \setminus j})$  is the joint belief. These definitions effectively remove the entropies for observed variables from the BFE evaluation. Note that although  $q_{a \setminus j}(s_{a \setminus j})$  is technically not a part of our belief set (7), it can be obtained by marginalization of  $q_a(s_a)$  (9b).

#### 5.2. Evaluation of Node-Local Free Energy for Deterministic Nodes

Another difficulty arises with the evaluation of the node-local free energy  $F[q_a]$  for factors of the form

$$f_a(s_a) = \delta(h_a(s_a)). \quad (73)$$

This type of node function reflects deterministic operations, e.g.,  $h(x, y, z) = z - x - y$  corresponds to the summation  $z = x + y$ . In this case, directly evaluating  $F[q_a]$  again leads to singularities.

There are (at least) two strategies available in the literature that resolve this issue. The first strategy “softens” the Dirac-delta by re-defining:

$$f_a(s_a) \triangleq \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{1}{2\epsilon} h_a(s_a)^2\right),$$

with  $0 < \epsilon \ll 1$  [17]. A drawback of this approach is that it may alter the model definition in a numerically unstable way, leading to a different inference solution and variational free energy than originally intended.

The second strategy combines the deterministic factor  $f_a$  with a neighboring stochastic factor  $f_b$  into a new *composite* factor  $f_c$ , by marginalizing over a shared variable  $s_j$ , leading to [60]

$$f_c(s_c) \triangleq \int \delta(h_a(s_a)) f_b(s_b) ds_j,$$

where  $s_c = \{s_a \cup s_b\} \setminus s_j$ . This procedure has drawbacks for models that involve many deterministic factors—namely, the convenient model modularity and resulting distributed compatibility are lost when large groups of factors are compacted in model-specific composite factors. We propose here a third strategy.

**Theorem 8.** Let  $f_a(s_a) = \delta(h_a(s_a))$ , with  $h_a(s_a) = s_j - g_a(s_{a \setminus j})$ , and node-local belief  $q_a(s_a) = q_{j|a}(s_j|s_{a \setminus j}) q_{a \setminus j}(s_{a \setminus j})$ . Then, the node-local free energy evaluates to

$$F[q_a, f_a] = \begin{cases} -H[q_{a \setminus j}] & \text{if } q_{j|a}(s_j|s_{a \setminus j}) = \delta(s_j - g_a(s_{a \setminus j})) \\ \infty & \text{otherwise.} \end{cases}$$

**Proof.** See Appendix D.16. □

An example that evaluates the node-local free energy for a non-trivial deterministic node can be found in Appendix C.

The equality node is a special case deterministic node, with a node function of the form (3). The argument of (Theorem 8) does not directly apply to this node. As the equality node function comprises two Dirac-delta functions, it can not be written in the form of Theorem 8. However, we can still reduce the node-local free energy contribution.

**Theorem 9.** Let  $f_a(s_a) = \delta(s_j - s_i) \delta(s_j - s_k)$ , with node-local belief  $q_a(s_a) = q_{ik|j}(s_i, s_k|s_j) q_j(s_j)$ . Then, the node-local free energy evaluates to

$$F[q_a, f_a] = \begin{cases} -H[q_j] & \text{if } q_{ik|j}(s_i, s_k|s_j) = \delta(s_j - s_i) \delta(s_j - s_k) \\ \infty & \text{otherwise.} \end{cases}$$

**Proof.** See Appendix D.17. □

### 5.3. Evaluating the Variational Free Energy

We propose here an algorithm that evaluates the BFE on a TFFG representation of a factorized model. The algorithm is based on the following results:

- The definitions for the computation of data-constrained entropies ensure that only variables with associated stochastic beliefs count towards the Bethe entropy. This makes the BFE evaluation consistent with Theorems 3 and 7, where the single-variable beliefs for observed variables are excluded from the BFE definition;
- We assume that a local mean-field factorization  $l(a)$  is available for each  $a \in \mathcal{V}$  (Section 4.1). If the mean-field factorization is not explicitly defined, we assume  $l(a) = \{a\}$  is the unfactored set;
- Deterministic nodes are accounted for by Theorem 8, which reduces the joint entropy to an entropy over the “inbound” edges. Although the belief over the “inbounds”  $q_{a \setminus j}(s_{a \setminus j})$  is not a term in the Bethe factorization (8), it can simply be obtained by marginalization of  $q_a(s_a)$ ;
- The equality node is a special case, where we let the node entropy discount the degree of the associated variable in the original model definition. While the BFE definition on a TFFG (7) does not explicitly account for edge degrees, this mechanism implicitly corrects for “double-counting” [17]. In this case, edge selection for counting is arbitrary, because all associated edges are (by definition) constrained to share the same belief (Section 2.1, Theorem 9).

The decomposition of (7) shows that the BFE can be computed by an iteration over the nodes and edges of the graph. As some contributions to the BFE might cancel each other, the algorithm first tracks counting numbers  $u_a$  for the average energies

$$U_a[q_a] = - \int q_a(s_a) \log f_a(s_a) ds_a,$$

and counting numbers  $h_k$  for the (joint) entropies

$$H[q_k] = - \int q_k(\mathbf{s}_k) \log q_k(\mathbf{s}_k) d\mathbf{s}_k,$$

which are ultimately combined and evaluated. We used an index  $k$  to indicate that the entropy computation may include not only the edges but a generic set of variables. We will give the definition of the set that  $k$  belongs to in Algorithm 1.

---

**Algorithm 1** Evaluation of the Bethe free energy on a Terminated Forney-style factor graph.

---

**given** a TFFG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

**given** a local mean-field factorization  $l(a)$  for all  $a \in \mathcal{V}$

**define**  $q_j(s_j) = \delta(s_j - \hat{s}_j) \Rightarrow H[q_j] \triangleq 0$   $\triangleright$  Ignore entropy of Dirac-delta constrained beliefs

**define**  $q_a(\mathbf{s}_a) = q_{a|j}(\mathbf{s}_{a \setminus j} | s_j) \delta(s_j - \hat{s}_j) \Rightarrow H[q_a] \triangleq H[q_{a \setminus j}]$   $\triangleright$  Reduce entropy of Dirac-delta constrained joint beliefs

**define**  $\mathcal{K} = \{a, a \setminus i, n, \text{ for all } a \in \mathcal{V}, i \in \mathcal{E}(a), n \in l(a)\}$  the set of (joint) belief indices

**initialize** counting numbers  $u_a = 0$  for all  $a \in \mathcal{V}$ ,  $h_k = 0$  for all  $k \in \mathcal{K}$

**for all** nodes  $a \in \mathcal{V}$  **do**

**if**  $a$  is a stochastic node **then**

$u_a += 1$

$\triangleright$  Count the average energy

**for all** clusters  $n \in l(a)$  **do**

$h_n += 1$

$\triangleright$  Count the (joint) cluster entropy

**end for**

**else if**  $a$  is an equality node **then**

    Select an edge  $j \in \mathcal{E}(a)$

$h_j += 1$

$\triangleright$  Count the variable entropy

**else**

$\triangleright$  Deterministic node  $a$

    Obtain the node function  $f_a(\mathbf{s}_a) = \delta(s_j - g_a(\mathbf{s}_{a \setminus j}))$

$h_{a \setminus j} += 1$

$\triangleright$  Count the (joint) entropy of the inbounds

**end if**

**end for**

**for all** edges  $i \in \mathcal{E}$  **do**

$h_i -= 1$

$\triangleright$  Discount the variable entropy

**end for**

$U = \sum_{a \in \mathcal{V}} u_a U_a[q_a]$

$H = \sum_{k \in \mathcal{K}} h_k H[q_k]$

**return**  $F = U - H$

---

## 6. Implementation of Algorithms and Simulations

We have developed a probabilistic programming toolbox *ForneyLab.jl* in the Julia language [61,62]. The majority of algorithms that are reviewed in Table 1 have been implemented in ForneyLab along with variety of demos (<https://github.com/biaslab/ForneyLab.jl/tree/master/demo>, accessed on 23 June 2021). ForneyLab is extendable and supports postulating new local constraints of the BFE for the creation of custom message passing-based inference algorithms.

In order to limit the length of this paper, we refer the reader to the demonstration folder of ForneyLab and to several of our previous papers with code. For instance, our previous work in [63] implemented a mean-field variational Laplace propagation for the hierarchical Gaussian filter (HGF) [64]. In the follow-up work [65], inference results improved by changing to structured factorization and moment-matching local constraints. In that case, modification of local constraints created a hybrid EP-VMP algorithm that better suited the model. Moreover, in [13], we formulated the idea of *chance constraints* in the form of violation probabilities leading to a new message passing algorithm that supports goal-directed behavior within the context of active inference. A similar line of reasoning led to improved inference procedures for auto-regressive models [66].

## 7. Related Work

Our work is inspired by the seminal work [17], which discusses the equivalence between the fixed points of the belief propagation algorithm [32] and the stationary points of the Bethe free energy. This equivalence is established through a Lagrangian formalism, which allows for the derivation of Generalized Belief Propagation (GBP) algorithms by introducing region-based graphs and the region-based (Kikuchi) free energy [16].

Region graph-based methods allows for overlapping clusters (Section 4.1) and thus offer a more generic message passing approach. The selection of appropriate regions (clusters), however, proves to be difficult, and the resulting algorithms may grow prohibitively complex. In this context, Ref. [67] addresses how to manipulate regions and manage the complexity of GBP algorithms. Furthermore, Ref. [68] also establishes a connection between GBP and expectation propagation (EP) by introducing structured region graphs.

The inspirational work of [15] derives message passing algorithms by minimization of  $\alpha$ -divergences. The stationary points of  $\alpha$ -divergences are obtained by a fixed point projection scheme. This projection scheme is reminiscent of the minimization scheme of the expectation propagation (EP) algorithm [18]. Compared to [15], our work focuses on a single divergence objective (namely, the VFE). The work of [12] derives the EP algorithm by manipulating the marginalization and factorization constraints of the Bethe free energy objective (see also Section 4.2.3). The EP algorithm is, however, not guaranteed to converge to a minimum of the associated divergence metric.

To address the convergence properties of the algorithms that are obtained by region graph methods, the outstanding work of [33] derives conditions on the region counting numbers that guarantee the convexity of the underlying objective. In general, however, the constrained Bethe free energy is not guaranteed to be convex and therefore the derived message passing updates are not guaranteed to converge.

## 8. Discussion

The key message in this paper is that a (variational) Bayesian model designer may tune the tractability-accuracy trade-off for evidence and posterior evaluation through constraint manipulation. It is interesting to note that the technique to derive message passing algorithms is always the same. We followed the recipe pioneered in [15] to derive a large variety of message passing algorithms solely through minimizing constrained Bethe free energy. This minimization leads to local fixed-point equations, which we can interpret as message passing updates on a (terminated) FFG. The presented lemmas showed how the constraints affect the Lagrangians locally. The presented theorems determined the stationary solutions of the Lagrangians and obtained the message passing equations. Thus, if a designer proposes a new set of constraints, then the first place to start is to analyze the effect on the Lagrangian. Once the effect of the constraint on the Lagrangian is known, then variational optimization may result in stationary solutions that can be obtained by a fixed-point iteration scheme.

In this paper, we selected the Forney-style factor graph framework to illustrate our ideas. FFGs are mathematically comparable to the more common bi-partite factor graphs that associate round nodes with variables and square nodes with factors [20]. Bi-partite

factor graphs require two distinct types of message updates (one leaving variable nodes and one leaving factor nodes), while message passing on a (T)FFG requires only a single type of message update [69]. The (T)FFG paradigm thus substantially simplifies the derivations and resulting message passing update equations.

The message passing update rules in this paper are presented without guarantees on convergence of the (local) minimization process. In practice, however, algorithm convergence can be easily checked by evaluating the BFE (Algorithm 1) after each belief update.

In future work, we plan on extending the treatment of constraints to formulate sampling-based algorithms such as importance sampling and Hamiltonian Monte Carlo in a message passing framework. While introducing SVMP, we have limited the discussion to local clusters that are not overlapping. We plan to extend variational algorithms to include local clusters that are overlapping without altering the underlying free-energy objective or the graph structure [16,67].

## 9. Conclusions

In this paper, we formulated a message-passing approach to probabilistic inference by identifying local stationary solutions of a constrained Bethe free energy objective (Sections 3 and 4). The proposed framework constructs a graph for the generative model and specifies local constraints for variational optimization in a local polytope. The constraints are then imposed on the variational objective by a Lagrangian construct. Unconstrained optimization of the Lagrangian then leads to local expressions of stationary points, which can be obtained by iterative execution of the resulting fixed point equations, which we identify with message passing updates.

Furthermore, we presented an approach to evaluate the BFE on a (terminated) Forney-style factor graph (Section 5). This procedure allows an algorithm designer to readily assess the performance of algorithms and models.

We have included detailed derivations of message passing updates (Appendix D) and hope that the presented formulation inspires the discovery of novel and customized message passing algorithms.

**Author Contributions:** Conceptualization: İ.Ş., T.v.d.L. and B.d.V.; methodology: İ.Ş. and T.v.d.L.; formal analysis, İ.Ş. and T.v.d.L.; investigation: İ.Ş., T.v.d.L. and B.d.V.; software, İ.Ş. and D.B.; validation, İ.Ş.; resources: İ.Ş., T.v.d.L. and B.d.V.; writing—original draft preparation: İ.Ş. and T.v.d.L.; writing—review and editing: T.v.d.L., D.B. and B.d.V.; visualizations: İ.Ş., T.v.d.L. and B.d.V.; supervision: T.v.d.L. and B.d.V.; project administration, B.d.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partly financed by GN Hearing A/S.

**Acknowledgments:** The authors would like to thank the BIASlab team members for many very interesting discussions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

BFE	Bethe Free Energy
BP	Belief Propagation
DC	Data Constraint
EM	Expectation Maximization
EP	Expectation Propagation
FFG	Forney-style Factor Graph
GBP	Generalized Belief Propagation
LP	Laplace Propagation
MFVLP	Mean-Field Variational Laplace
MFVMP	Mean-Field Variational Message Passing



NLE	Negative Log-Evidence
TFFG	Terminated Forney-style Factor Graph
VFE	Variational Free Energy
VMP	Variational Message Passing
SVMP	Structured Variational Message Passing
SP	Sum-Product

## Appendix A. Free Energy Minimization by Variational Inference

In this section, we present a pedagogical example of inductive inference. After we establish an intuition, we apply the same principles to a more general context in the further sections. We follow Caticha [42,70], who showed that a constrained free energy functional can be interpreted as a principled objective measure for inductive reasoning, see also [71,72]. The calculus of variations offers a principled method for optimizing this free energy functional.

In this section, we assume an example model

$$f(\mathbf{y}, \theta) = f_{\mathbf{y}}(\mathbf{y}, \theta) f_{\theta}(\theta), \quad (\text{A1})$$

with observed variables  $\mathbf{y}$  and a single parameter  $\theta$ .

We define the (variational) free energy (VFE) as

$$F[q, f] = \iint q(\mathbf{y}, \theta) \log \frac{q(\mathbf{y}, \theta)}{f(\mathbf{y}, \theta)} d\mathbf{y} d\theta. \quad (\text{A2})$$

The goal is to find a posterior

$$q^* = \arg \min_{q \in \mathcal{Q}} F[q, f] \quad (\text{A3})$$

that minimizes the free energy subject to some pre-specified constraints. These constraints may include form or factorization constraints on  $q$  (to be discussed later) or relate to observations of a signal  $\mathbf{y}$ .

As an example, assume that we obtained some measurements  $\mathbf{y} = \hat{\mathbf{y}}$  and wish to obtain a posterior marginal belief  $q^*(\theta)$  over the parameter. We can then incorporate the data in the form of a data constraint

$$\int q(\mathbf{y}, \theta) d\theta = \delta(\mathbf{y} - \hat{\mathbf{y}}), \quad (\text{A4})$$

where  $\delta$  defines a Dirac-delta. The *constrained* free energy can be rewritten by including Lagrange multipliers as

$$L[q, f] = F[q, f] + \gamma \left( \iint q(\mathbf{y}, \theta) d\mathbf{y} d\theta - 1 \right) + \int \lambda(\mathbf{y}) \left( \int q(\mathbf{y}, \theta) d\theta - \delta(\mathbf{y} - \hat{\mathbf{y}}) \right) d\mathbf{y}, \quad (\text{A5})$$

where the first term specifies the (to be minimized) free energy objective, the second term a normalization constraint, and the third term the data constraint. Optimization of (A5) can be performed using variational calculus.

Variational calculus considers the impact of a variation in  $q(\mathbf{y}, \theta)$  on the Lagrangian  $L[q, f]$ . We define the variation as

$$\delta q(\mathbf{y}, \theta) \triangleq \epsilon \phi(\mathbf{y}, \theta),$$

where  $\epsilon \rightarrow 0$ , and  $\phi$  is a continuous and differentiable “test” function. The fundamental theorem of variational calculus states that the stationary solutions  $q^*$  are obtained by

setting  $\delta L/\delta q = 0$ , where the functional derivative  $\delta L/\delta q$  is implicitly defined by Appendix D in [2]:

$$\left. \frac{dL[q + \epsilon\phi, f]}{d\epsilon} \right|_{\epsilon=0} = \iint \frac{\delta L}{\delta q}(\mathbf{y}, \theta) \phi(\mathbf{y}, \theta) d\mathbf{y} d\theta. \tag{A6}$$

Equation (A6) provides a way to derive the functional derivative through ordinary differentiation. For example, we take the Lagrangian defined by (A5) and work out the left hand side of (A6):

$$\left. \frac{dL[q + \epsilon\phi, f]}{d\epsilon} \right|_{\epsilon=0} = \left. \frac{dF[q + \epsilon\phi, f]}{d\epsilon} \right|_{\epsilon=0} + \left. \frac{d}{d\epsilon} \gamma \iint (q + \epsilon\phi) d\mathbf{y} d\theta \right|_{\epsilon=0} + \left. \frac{d}{d\epsilon} \int \lambda(\mathbf{y}) \int (q + \epsilon\phi) d\theta d\mathbf{y} \right|_{\epsilon=0} \tag{A7a}$$

$$= \iint \left. \frac{d}{d\epsilon} \left( (q + \epsilon\phi) \log \frac{(q + \epsilon\phi)}{f} \right) \right|_{\epsilon=0} d\mathbf{y} d\theta + \gamma \iint \left. \frac{d}{d\epsilon} (q + \epsilon\phi) \right|_{\epsilon=0} d\mathbf{y} d\theta + \int \lambda(\mathbf{y}) \int \left. \frac{d}{d\epsilon} (q + \epsilon\phi) \right|_{\epsilon=0} d\theta d\mathbf{y} \tag{A7b}$$

$$= \iint \underbrace{\left[ \log \frac{q(\mathbf{y}, \theta)}{f(\mathbf{y}, \theta)} + 1 + \gamma + \lambda(\mathbf{y}) \right]}_{\delta L[q, f]/\delta q} \phi(\mathbf{y}, \theta) d\mathbf{y} d\theta. \tag{A7c}$$

Note that, since (A7c) has been written in similar form as (A6), it is easy to identify the functional derivative. This procedure is one of many ways to obtain the functional derivatives [73].

Setting  $\delta L[q, f]/\delta q = 0$  we find the stationary solution as

$$q^*(\mathbf{y}, \theta) = \exp(-1 - \gamma - \lambda(\mathbf{y})) f(\mathbf{y}, \theta) \tag{A8a}$$

$$= \frac{1}{Z} \exp(-\lambda(\mathbf{y})) f(\mathbf{y}, \theta), \tag{A8b}$$

with  $Z = \iint \exp(-\lambda(\mathbf{y})) f(\mathbf{y}, \theta) d\mathbf{y} d\theta = \exp(\gamma + 1)$ . In order to determine the Lagrange multipliers  $\gamma$  and  $\lambda(\mathbf{y})$ , we must substitute the stationary solution (A8b) back into the constraints. The normalization constraint evaluates to

$$\frac{1}{Z} \iint \exp(-\lambda(\mathbf{y})) f(\mathbf{y}, \theta) d\mathbf{y} d\theta = 1. \tag{A9}$$

We find that (A9) is always satisfied since  $Z = \iint \exp(-\lambda(\mathbf{y})) f(\mathbf{y}, \theta) d\mathbf{y} d\theta$  by definition. Note, however, that the computation of the normalization constant still depends on the undetermined Lagrange multiplier  $\lambda(\mathbf{y})$ .

The data constraint evaluates to

$$\int q^*(\mathbf{y}, \theta) d\theta = \frac{1}{Z} \exp(-\lambda(\mathbf{y})) \int f(\mathbf{y}, \theta) d\theta = \delta(\mathbf{y} - \hat{\mathbf{y}}) \tag{A10}$$

which can be rewritten as

$$\frac{\exp(-\lambda(\mathbf{y}))}{Z} = \frac{\delta(\mathbf{y} - \hat{\mathbf{y}})}{\int f(\mathbf{y}, \theta) d\theta}. \tag{A11}$$

Equation (A11) shows that  $\lambda(\mathbf{y})$  can satisfy this constraint only if it is proportional to  $\delta(\mathbf{y} - \hat{\mathbf{y}})$ . Indeed, substitution of (A11) into (A8b) gives

$$q^*(\mathbf{y}, \theta) = \frac{f(\mathbf{y}, \theta)}{\int f(\mathbf{y}, \theta) d\theta} \delta(\mathbf{y} - \hat{\mathbf{y}}),$$

and the posterior for the parameters evaluates to

$$\begin{aligned} q^*(\theta) &= \int q^*(\mathbf{y}, \theta) d\mathbf{y} \\ &= \int \frac{f(\mathbf{y}, \theta)}{\int f(\mathbf{y}, \theta) d\theta} \delta(\mathbf{y} - \hat{\mathbf{y}}) d\mathbf{y} \\ &= \frac{f(\hat{\mathbf{y}}, \theta)}{\int f(\hat{\mathbf{y}}, \theta) d\theta} \\ &= \frac{f_{\mathbf{y}}(\hat{\mathbf{y}}, \theta) f_{\theta}(\theta)}{\int f_{\mathbf{y}}(\hat{\mathbf{y}}, \theta) f_{\theta}(\theta) d\theta}, \end{aligned}$$

which we recognize as the Bayes rule.

Note that the Bayes rule was derived here as a special case of constrained variational free energy minimization when data constraints are present. This derivation of the Bayes rule seems unnecessarily tedious but the value of this approach to inductive inference is that the same principle applies when other (not data) constraints on  $q$  are present.

## Appendix B. Lagrangian Optimization and the Dual Problem

With the addition of Lagrange multipliers to the Bethe functional, the resulting Lagrangian depends both on the variational distribution  $q(s)$  and the Lagrange multipliers  $\Psi(s)$ . Formally, the introduction of the Lagrange multipliers allows us to rewrite the constrained optimization on the local polytope as an unconstrained optimization. We follow [33], and write

$$\min_{q \in \mathcal{L}(\mathcal{G})} F[q] = \min_q \max_{\Psi} L[q, \Psi].$$

Weak duality [74] (Chapter 5) then states that

$$\min_q \max_{\Psi} L[q, \Psi] \geq \max_{\Psi} \min_q L[q, \Psi].$$

The minimization with respect to  $q$  then yields a solution that depends on the Lagrange multipliers, as

$$q^*(s; \Psi) = \arg \min_q L[q, \Psi].$$

For any given  $q$  the Lagrangian is concave in  $\Psi$ . Therefore, substituting  $q^*$  in the Lagrangian, the maximization over  $L[q^*, \Psi]$  yields the unique solution

$$\Psi^*(s) = \arg \max_{\Psi} L[q^*, \Psi].$$

Stationary solutions are then given by

$$q^*(s; \Psi^*) = \arg \min_{q \in \mathcal{L}(\mathcal{G})} F[q].$$

In the current paper, we consider factorized  $q$ 's (e.g., (8)), and consider variations with respect to the individual factors. We then need to show that the combined stationary points of the individual factors also constitute a stationary point of the total objective.

Consider a Lagrangian having multiple arguments, i.e.,

$$L[q] = L[q_1, \dots, q_n, \dots, q_N] \tag{A12}$$

$$\mathbf{q} \triangleq [q_1, \dots, q_N]^T. \tag{A13}$$

We want to determine the first total variation of the Lagrangian given by

$$\delta L = L[\mathbf{q} + \epsilon \boldsymbol{\phi}] - L[\mathbf{q}] \quad (\text{A14})$$

$$\boldsymbol{\phi}(\mathbf{s}) \triangleq [\phi_1(\mathbf{s}), \dots, \phi_N(\mathbf{s})]^\top. \quad (\text{A15})$$

By a Taylor series expansion on  $\epsilon$  we obtain [73] (A.14) and [75] (Equation (23.2))

$$L[\mathbf{q} + \epsilon \boldsymbol{\phi}] - L[\mathbf{q}] = \sum_{k=1}^K \frac{1}{k!} \frac{d}{d\epsilon^k} \left( L^k[\mathbf{q} + \epsilon \boldsymbol{\phi}] \right) \epsilon^k + \mathcal{O}(\epsilon^{K+1}). \quad (\text{A16})$$

Omitting all terms higher than the first order, we obtain the first variation as

$$\delta L = \frac{d}{d\epsilon} (L[\mathbf{q} + \epsilon \boldsymbol{\phi}]) \epsilon. \quad (\text{A17})$$

Rearranging the terms and letting  $\epsilon$  vanish, we obtain the following expression

$$\lim_{\epsilon \rightarrow 0} \frac{\delta L}{\epsilon} = \frac{d}{d\epsilon} (L[\mathbf{q} + \epsilon \boldsymbol{\phi}]) \Big|_{\epsilon=0}. \quad (\text{A18})$$

Let us assume that the Frechet derivative exists [73] such that we can obtain the following integral representation (It should be noted that this integral expression is not always possible for a generic Lagrangian. That is why we need to assume that the Frechet derivative exists)

$$\frac{d}{d\epsilon} (L[\mathbf{q} + \epsilon \boldsymbol{\phi}]) \Big|_{\epsilon=0} = \int \boldsymbol{\phi}(\mathbf{s})^\top \frac{\delta L}{\delta \mathbf{q}} d\mathbf{s} \quad (\text{A19})$$

where  $\frac{\delta L}{\delta \mathbf{q}}$  is the variational derivative

$$\frac{\delta L}{\delta \mathbf{q}} = \left[ \frac{\delta L}{\delta q_1}, \dots, \frac{\delta L}{\delta q_N} \right]^\top \quad (\text{A20})$$

$$\delta q_n = \epsilon \phi_n(\mathbf{s}). \quad (\text{A21})$$

This means that (A19) can be written as [75] (Equation (22.5)) (Here, we use a more generic Lagrangian and our notation is different than in [75]; however the expression is motivated again by a Taylor series expansion on  $\epsilon$ )

$$\lim_{\epsilon \rightarrow 0} \frac{\delta L}{\epsilon} = \frac{d}{d\epsilon} (L[\mathbf{q} + \epsilon \boldsymbol{\phi}]) \Big|_{\epsilon=0} = \sum_n \int \boldsymbol{\phi}(\mathbf{s}) \frac{\delta L}{\delta q_n} d\mathbf{s}. \quad (\text{A22})$$

Fundamental theorem of variational calculus states that in order for a point to be stationary, the first variation needs to vanish. In order for the first variation to vanish, it is sufficient to have vanishing of the variational derivatives

$$\frac{\delta L}{\delta q_n} = 0 \text{ for every } n = 1, \dots, N. \quad (\text{A23})$$

Vanishing of individual variational derivatives will mean that that the local stationary points will also correspond to a global stationary point.

### Appendix C. Local Free Energy Example for a Deterministic Node

Theorem 8 tells us how to evaluate the node-local free energy for a deterministic node. As an example, consider the node function  $f_a(y, x) = \delta(y - \text{sgn}(x))$ , with  $y \in \{-1, 1\}$  and  $x \in \mathbb{R}$  as depicted in Figure A1.

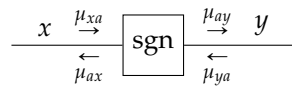


Figure A1. Messages around a “sign” node.

Interestingly, there is information loss in this node because the “sign” mapping is not bijective. Given an incoming Bernoulli distributed message  $\mu_{ya}(y) = \text{Ber}(y|p)$ , the backward outgoing message is derived as

$$\begin{aligned} \mu_{ax}(x) &= \int \mu_{ya}(y) \delta(y - \text{sgn}(x)) dy \\ &= \begin{cases} p & \text{if } x \geq 0 \\ 1 - p & \text{if } x < 0. \end{cases} \end{aligned}$$

Given a Gaussian distributed incoming message  $\mu_{xa}(x) = \mathcal{N}(x|m, \vartheta)$ , the resulting belief then becomes

$$\begin{aligned} q_x(x) &= \frac{\mu_{xa}(x) \mu_{ax}(x)}{\int \mu_{xa}(x) \mu_{ax}(x) dx} \\ &= \begin{cases} \frac{p}{p + \Phi - 2p\Phi} \mathcal{N}(x|m, \vartheta) & \text{if } x \geq 0 \\ \frac{1-p}{p + \Phi - 2p\Phi} \mathcal{N}(x|m, \vartheta) & \text{if } x < 0, \end{cases} \end{aligned}$$

with  $\Phi = \int_{-\infty}^0 \mathcal{N}(x|m, \vartheta) dx$ . We define a truncated Gaussian distribution as

$$\mathcal{T}(x|m, \vartheta, a, b) = \begin{cases} \frac{1}{\Phi(a,b;m,\vartheta)} \mathcal{N}(x|m, \vartheta) & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise,} \end{cases}$$

with  $\Phi(a, b; m, \vartheta) = \int_a^b \mathcal{N}(x|m, \vartheta) dx$ . This leads to

$$q_x(x) = \underbrace{\frac{p(1-\Phi)}{p + \Phi - 2p\Phi}}_K \mathcal{T}(x|m, \vartheta, -\infty, 0) + \underbrace{\frac{(1-p)\Phi}{p + \Phi - 2p\Phi}}_{1-K} \mathcal{T}(x|m, \vartheta, 0, \infty),$$

as a truncated Gaussian mixture.

The node-local free energy then evaluates to

$$\begin{aligned} F[q_a, f_a] &= -H[q_x] \\ &= \int_{-\infty}^0 q_x(x) \log q_x(x) dx + \int_0^{\infty} q_x(x) \log q_x(x) dx \\ &= -KH[\mathcal{T}(m, \vartheta, -\infty, 0)] + K \log K - (1 - K)H[\mathcal{T}(m, \vartheta, 0, \infty)] + (1 - K) \log(1 - K) \\ &= -KH[\mathcal{T}(m, \vartheta, -\infty, 0)] - (1 - K)H[\mathcal{T}(m, \vartheta, 0, \infty)] - H[\text{Ber}(K)], \end{aligned}$$

as a weighted sum of entropies, which can be computed analytically.

### Appendix D. Proofs

#### Appendix D.1. Proof of Lemma 1

**Proof.** We apply the variation  $\epsilon\phi_b$  to  $q_b$  and, as discussed in Appendix A, we can identify the functional derivative  $\delta L_b/\delta q_b$  through ordinary differentiation as

$$\left. \frac{dL_b[q_b + \epsilon\phi_b, f_b]}{d\epsilon} \right|_{\epsilon=0} = \int \left( \overbrace{\log \frac{q_b(s_b)}{f_b(s_b)} + 1 + \psi_b - \sum_{i \in \mathcal{E}(b)} \lambda_{ib}(s_i)}^{\delta L_b/\delta q_b} \right) \phi_b(s_b) ds_b.$$

Setting the functional derivative to zero and identifying

$$\mu_{ib}(s_i) = \exp(\lambda_{ib}(s_i)) \tag{A24}$$

$$\psi_b = \log \int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i) d\mathbf{s}_b - 1 \tag{A25}$$

yields the stationary solutions (18) in terms of Lagrange multipliers that are to be determined.  $\square$

Appendix D.2. Proof of Lemma 2

**Proof.** We follow the same procedure as in Appendix D.1, where we apply a variation  $\epsilon\phi_j$  to  $q_j$  (instead of  $q_b$ ), and identify the functional derivative  $\delta L_j / \delta q_j$  through

$$\left. \frac{dL_j[q_j + \epsilon\phi_j]}{d\epsilon} \right|_{\epsilon=0} = \int \left( \overbrace{-\log q_j(s_j) - 1 + \psi_j + \sum_{a \in \mathcal{V}(j)} \lambda_{ja}(s_j)}^{\delta L_j / \delta q_j} \right) \phi_j(s_j) ds_j.$$

As the TFFG is terminated, each edge has 2 degrees and the node-induced edge set has only 2 factors, which we denote by  $f_b$  and  $f_c$ . Then, setting the functional derivative to zero and identifying

$$\mu_{ja}(s_j) = \exp(\lambda_{ja}(s_j)) \tag{A26}$$

$$\psi_j = -\log \int \mu_{jb}(s_j) \mu_{jc}(s_j) ds_j + 1 \tag{A27}$$

yields the stationary solution of (20) in terms of the Lagrange multipliers.  $\square$

Appendix D.3. Proof of Theorem 1

**Proof.** The local polytope of (14) constructs the Lagrangians of (17) and (19). Substituting the stationary solutions from Lemmas 1 and 2 in the marginalization constraint,

$$q_j(s_j) = \int q_b(\mathbf{s}_b) d\mathbf{s}_{b \setminus j},$$

we obtain the following relation

$$\frac{\mu_{jb}(s_j) \mu_{jc}(s_j)}{Z_j} = \frac{1}{Z_b} \int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i) d\mathbf{s}_{b \setminus j},$$

where we defined the following normalization constants to ensure that the computed marginals are normalized:

$$Z_j = \int \mu_{jb}(s_j) \mu_{jc}(s_j) ds_j$$

$$Z_b = \int f_b(\mathbf{s}_b) \prod_{i \in \mathcal{E}(b)} \mu_{ib}(s_i) d\mathbf{s}_b.$$

Extracting  $\mu_{jb}$  from the integral

$$\begin{aligned} \frac{\mu_{jb}(s_j)\mu_{jc}(s_j)}{Z_j} &= \frac{\mu_{jb}(s_j)}{Z_b} \int f_b(\mathbf{s}_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}(s_i) \, d\mathbf{s}_{b \setminus j}, \\ \mu_{jc}(s_j) &= \frac{Z_j}{Z_b} \int f_b(\mathbf{s}_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}(s_i) \, d\mathbf{s}_{b \setminus j} \end{aligned} \tag{A28}$$

and cancelling  $\mu_{jb}$  on both sides then yields the condition on the functional form of the message  $\mu_{jc}$ .

We now need to show that the fixed points of (22) satisfy (A28). Let us assume that the fixed points exist, such that  $\mu_{jc}^{(k)} = \mu_{jc}^{(k+1)}$  for some  $k$ . Then, we want to show that at the fixed points the following equality holds:

$$\mu_{jc}^{(k)}(s_j) = \frac{Z_j^{(k)}}{Z_b^{(k)}} \int f_b(\mathbf{s}_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}^{(k)}(s_i) \, d\mathbf{s}_{b \setminus j}.$$

Substituting (22), we need to show that

$$\mu_{jc}^{(k)}(s_j) = \frac{Z_j^{(k)}}{Z_b^{(k)}} \mu_{jc}^{(k+1)}(s_j).$$

Since  $\mu_{jc}^{(k)} = \mu_{jc}^{(k+1)}$ , we can rearrange

$$\mu_{jc}^{(k)} \left( 1 - \frac{Z_j^{(k)}}{Z_b^{(k)}} \right) = 0.$$

From  $Z_b$ , we obtain

$$\begin{aligned} Z_b^{(k)} &= \int \mu_{jb}^{(k)}(s_j) \int f_b(\mathbf{s}_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}^{(k)}(s_i) \, d\mathbf{s}_{b \setminus j} \, ds_j \\ &= \int \mu_{jb}^{(k)}(s_j) \mu_{jc}^{(k+1)}(s_j) \, ds_j \\ &= \int \mu_{jb}^{(k)}(s_j) \mu_{jc}^{(k)}(s_j) \, ds_j \\ &= Z_j^{(k)}, \end{aligned}$$

which implies that the fixed points satisfy the desired condition. This proves that the stationary solutions to the BFE within the local polytope can be obtained as fixed points of the sum-product update equations.  $\square$

Appendix D.4. Proof of Lemma 3

**Proof.** Substituting the definition of (32), we can re-write the second term of Lagrangian (30) as

$$\begin{aligned} \int \left\{ \prod_{n \in l(b)} q_b^m(\mathbf{s}_b^m) \right\} \log f_b(\mathbf{s}_b) \, d\mathbf{s}_b &= \int q_b^m(\mathbf{s}_b^m) \left( \int \left\{ \prod_{\substack{n \in l(b) \\ n \neq m}} q_b^n(\mathbf{s}_b^n) \right\} \log f_b(\mathbf{s}_b) \, d\mathbf{s}_{b \setminus m}^m \right) \, d\mathbf{s}_b^m \\ &= \int q_b^m(\mathbf{s}_b^m) \log \tilde{f}_b^m(\mathbf{s}_b^m) \, d\mathbf{s}_b^m. \end{aligned}$$



We apply the variation  $\epsilon\phi_b^m$  to  $q_b^m$  and identify the functional derivative  $\delta L_b^m/\delta q_b^m$ , as

$$\left. \frac{dL_b^m[q_b^m + \epsilon\phi_b^m]}{d\epsilon} \right|_{\epsilon=0} = \int \left( \overbrace{\log \frac{q_b^m(\mathbf{s}_b^m)}{f_b^m(\mathbf{s}_b^m)} + 1 + \psi_b^m - \sum_{i \in m} \lambda_{ib}(s_i)}^{\delta L_b^m/\delta q_b^m} \right) \phi_b^m(\mathbf{s}_b^m) d\mathbf{s}_b^m,$$

whose functional form we recognize from Appendix D.1. Setting the functional derivative to zero and again identifying  $\mu_{ib}(s_i) = \exp \lambda_{ib}(s_i)$ , yields the stationary solutions of (31).  $\square$

Appendix D.5. Proof of Theorem 2

**Proof.** The local polytope of (33) constructs the Lagrangians  $L_b^m$  and  $L_j$  as (30) and (19), respectively. We substitute the stationary solutions of Lemmas 2 and 3 in the local marginalization constraint (29b), which yields

$$q_j(s_j) = \int q_b^m(\mathbf{s}_b^m) d\mathbf{s}_{b \setminus j}^m.$$

Following the structure of the proof in Appendix D.3, we obtain the following condition for the stationary solutions in terms of messages:

$$\begin{aligned} \frac{\mu_{jb}(s_j)\mu_{jc}(s_j)}{Z_j} &= \frac{\mu_{jb}(s_j)}{Z_b^m} \int \tilde{f}_b^m(\mathbf{s}_b^m) \prod_{\substack{i \in m \\ i \neq j}} \mu_{ib}(s_i) d\mathbf{s}_{b \setminus j}^m \\ \frac{\mu_{jc}(s_j)}{Z_j} &= \frac{1}{Z_b^m} \int \tilde{f}_b^m(\mathbf{s}_b^m) \prod_{\substack{i \in m \\ i \neq j}} \mu_{ib}(s_i) d\mathbf{s}_{b \setminus j}^m. \end{aligned} \tag{A29}$$

Now we want to show that the fixed points of the message updates (36) satisfy (A29). Let us assume that the fixed points exists for some  $k$  such that  $\mu_{jc}^{(k+1)} = \mu_{jc}^{(k)}$ . Then, we will show that the fixed points satisfy

$$\frac{\mu_{jc}^{(k)}(s_j)}{Z_j^{(k)}} = \frac{1}{Z_b^{m,(k)}} \int \tilde{f}_b^{m,(k)}(\mathbf{s}_b^m) \prod_{\substack{i \in m \\ i \neq j}} \mu_{ib}^{(k)}(s_i) d\mathbf{s}_{b \setminus j}^m. \tag{A30}$$

Similar to Appendix D.3, it will suffice to show that  $Z_b^{m,(k)} = Z_j^{(k)}$  at the fixed points. Arranging the order of integration in normalization constant  $Z_b^{m,(k)}$ , we obtain

$$\begin{aligned} Z_b^{m,(k)} &= \int \mu_{jb}^{(k)}(s_j) \int \tilde{f}_b^{m,(k)}(\mathbf{s}_b^m) \prod_{\substack{i \in m \\ i \neq j}} \mu_{ib}^{(k)}(s_i) d\mathbf{s}_{b \setminus j}^m ds_j \\ &= \int \mu_{jb}^{(k)}(s_j) \mu_{jc}^{(k+1)}(s_j) ds_j \\ &= \int \mu_{jb}^{(k)}(s_j) \mu_{jc}^{(k)}(s_j) ds_j \\ &= Z_j^{(k)}. \end{aligned}$$

By the same line of reasoning as in Appendix D.3, this shows that the fixed points of the message updates (36) leads to stationary distributions of the Bethe free energy with structured factorization constraints.  $\square$

Appendix D.6. Proof of Corollary 1

**Proof.** For a fully factorized local variational distribution (41), the augmented node function  $\tilde{f}_b^m(\mathbf{s}_b^m)$  of (32) reduces to

$$\tilde{f}_j(s_j) = \exp \left( \int \left\{ \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} q_i(s_i) \right\} \log f_b(\mathbf{s}_b) \, d\mathbf{s}_{b \setminus j} \right). \tag{A31}$$

The message of (36) then reduces to

$$\mu_{jc}(s_j) = \tilde{f}_j(s_j),$$

which, after substitution, recovers (43).  $\square$

Appendix D.7. Proof of Lemma 4

**Proof.** When we apply the variation  $\epsilon\phi_b$  to  $q_b$  and identify the functional derivative  $\delta L_b / \delta q_b$ , we recover the result from Appendix D.1, which leads to a solution of the form (47).  $\square$

Appendix D.8. Proof of Theorem 3

**Proof.** We construct the Lagrangian of (46), which by Lemma 4 leads to a solution of the form (47). Substituting this solution in the constraint of (45) leads to

$$\left[ \int f_b(\mathbf{s}_b) \overbrace{\prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}(s_i)}^{\mu_{bj}(s_j)} \, d\mathbf{s}_{b \setminus j} \right] \mu_{jb}(s_j) = \delta(s_j - \hat{s}_j). \tag{A32}$$

This equation is then satisfied by (50), which proves the theorem.  $\square$

Appendix D.9. Proof of Lemma 5

**Proof.** The proof follows directly from Appendix D.1, with  $\tilde{f}_b(\mathbf{s}_b; \hat{\mathbf{s}}_b)$  substituted for  $f_b(\mathbf{s}_b)$ .  $\square$

Appendix D.10. Proof of Theorem 4

**Proof.** Given the result of Lemma 5, the proof follows Appendix D.3, where Laplace propagation chooses the expansion point to be the fixed point  $\hat{\mathbf{s}}_b = \arg \max \log q_b(\mathbf{s}_b)$ .

For all second-order fixed points of the Laplace iterations, it holds that  $\hat{\mathbf{s}}_b$  is a fixed point if and only if it is a local optimum of  $q_b$ . The proof is then concluded by Lemma 1 in [76].  $\square$

Appendix D.11. Proof of Lemma 6

**Proof.** We note that the Lagrange multiplier  $\eta_{jb}$  does not depend on  $s_j$  because the expectation removes all the functional dependencies on  $s_j$ . Furthermore, the expectations of  $T_j(s_j)$  have the same dimension as the function  $T_j(s_j)$ . This means that the dimension of  $\eta_{jb}$  needs to be compatible with that of  $T_j(s_j)$  so that we can write the constraint as an inner product.

We apply the variation  $\epsilon\phi_b$  to  $q_b$  and identify the functional derivative  $\delta L_b / \delta q_b$ , as

$$\left. \frac{dL_b[q_b + \epsilon\phi_b, f_b]}{d\epsilon} \right|_{\epsilon=0} = \int \left( \overbrace{\log \frac{q_b(\mathbf{s}_b)}{f_b(\mathbf{s}_b)} + 1 + \psi_b - \sum_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \lambda_{ib}(s_i) - \eta_{jb}^\top T_j(s_j)}^{\delta L_b / \delta q_b} \right) \phi_b(\mathbf{s}_b) \, d\mathbf{s}_b.$$

Setting the functional derivative to zero and identifying  $\mu_{ib}(s_i) = \exp \lambda_{ib}(s_i)$  for  $i \neq j$  and identifying  $\mu_{jb}(s_j) = \exp(\eta_{jb}^\top T_j(s_j))$  yields the functional form of the stationary solution as (62).  $\square$

Appendix D.12. Proof of Lemma 7

**Proof.** We follow a similar procedure as in Appendix D.11 and apply the variation  $\epsilon\phi_j$  to  $q_j$ , which identifies the functional derivative  $\delta L_j/\delta q_j$ , as

$$\frac{dL[q_j + \epsilon\phi_j]}{d\epsilon} \Big|_{\epsilon=0} = \int \left( \overbrace{-\log q_j(s_j) - 1 + \psi_j + \sum_{a \in \mathcal{V}(j)} \eta_{ja}^\top T_j(s_j)}^{\delta L_j/\delta q_j} \right) \phi_j(s_j) ds_j.$$

Setting the functional derivative to zero and following the same procedure as in Appendix D.2 yields (64).  $\square$

Appendix D.13. Proof of Theorem 5

**Proof.** By substituting the stationary solutions given by Lemmas 6 and 7 into the moment-matching constraint (59), we obtain the following condition:

$$\begin{aligned} \int T_j(s_j) q_j(s_j) ds_j &= \int T_j(s_j) q_b(s_b) ds_b \\ \frac{1}{Z_j} \int T_j(s_j) \exp([\eta_{jb} + \eta_{jc}]^\top T_j(s_j)) ds_j &= \frac{1}{\tilde{Z}_j} \int T_j(s_j) \overbrace{\exp(\eta_{jb}^\top T_j(s_j))}^{\mu_{jb}(s_j)} \left[ \int \overbrace{f_b(s_b) \prod_{\substack{i \in \mathcal{E}(b) \\ i \neq j}} \mu_{ib}(s_i)}^{\tilde{\mu}_{jc}(s_j)} ds_{b \setminus j} \right] ds_j \\ &= \int T_j(s_j) \tilde{q}_j(s_j) ds_j, \end{aligned}$$

where we recognize the sum-product message  $\tilde{\mu}_{jc}(s_j)$ , which we multiply by the incoming exponential family message  $\mu_{jb}(s_j)$  and normalize to obtain  $\tilde{q}_j(s_j)$ . By defining  $\eta_j = \eta_{jb} + \eta_{jc}$ , normalization constants are given by

$$\begin{aligned} Z_j(\eta_j) &= \int \exp(\eta_j^\top T_j(s_j)) ds_j \\ \tilde{Z}_j &= \int \exp(\eta_{jb}^\top T_j(s_j)) \tilde{\mu}_{jc}(s_j) ds_j. \end{aligned}$$

Computing the moments allows us to determine the exponential family parameter by solving the following equation [24] (Proposition 3.1)

$$\nabla_{\eta_j} \log Z_j(\eta_j) = \int \tilde{q}_j(s_j) T_j(s_j) ds_j.$$

Suppose you obtain a solution to this equation denoted by  $\tilde{\eta}_j$ , this allows us to approximate the sum-product message  $\tilde{\mu}_{jc}(s_j)$  by an exponential family message whose parameter is given by

$$\eta_{jc} = \tilde{\eta}_j - \eta_{jb}.$$

Now let us assume that the fixed points of the sum-product iterations  $\tilde{\mu}_{jc}^{(k)}(s_j) = \tilde{\mu}_{jc}^{(k+1)}(s_j)$  and the incoming exponential family messages  $\mu_{jb}^{(k)}(s_j) = \mu_{jb}^{(k+1)}(s_j)$  exist for some  $k$ . Then, we need to show that the existence of these fixed points implies the existence of the fixed points of  $\mu_{jc}^{(k+1)} = \mu_{jc}^{(k)}$ .

By moment-matching, we have

$$\begin{aligned} \eta_{jc}^{(k+1)} &= \tilde{\eta}_j^{(k+1)} - \eta_{jb}^{(k+1)} \\ &= \tilde{\eta}_j^{(k)} - \eta_{jb}^{(k)} \\ &= \eta_{jc}^{(k)}, \end{aligned}$$

which proves the existence of the fixed point of  $\mu_{jc}$  if  $\tilde{\mu}_{jc}$  and  $\mu_{jb}(s_j)$  have fixed points.  $\square$

Appendix D.14. Proof of Theorem 6

**Proof.** The proof follows directly from substituting the Laplace-approximated factor-function (53) in the naive mean-field result of Corollary 1.  $\square$

Appendix D.15. Proof of Theorem 7

**Proof.** In order to obtain the optimal parameter value  $\theta_j^*$ , we view the free energy as a function of  $\theta_j$ . As there are two node-local free energies that depend upon  $\theta_j$ , this leads to

$$\begin{aligned} \theta_j^* &= \arg \min_{\theta_j} \left( F[q_b, f_b; \theta_j] + F[q_c, f_c; \theta_j] \right) \\ &= \arg \max_{\theta_j} \left( \int \left\{ \delta(s_j - \theta_j) \prod_{\substack{n \in I(b) \\ n \neq m}} q_b^n(s_b^n) \right\} \log f_b(s_b) ds_b + \int \left\{ \delta(s_j - \theta_j) \prod_{\substack{n \in I(c) \\ n \neq m}} q_c^n(s_c^n) \right\} \log f_c(s_c) ds_c \right) \\ &= \arg \max_{\theta_j} \left( \int \left\{ \prod_{\substack{n \in I(b) \\ n \neq m}} q_b^n(s_b^n) \right\} \log f_b(s_{b \setminus j}, \theta_j) ds_{b \setminus j} + \int \left\{ \prod_{\substack{n \in I(c) \\ n \neq m}} q_c^n(s_c^n) \right\} \log f_c(s_{c \setminus j}, \theta_j) ds_{c \setminus j} \right) \\ &= \arg \max_{s_j} \left( \log \mu_{bj}(s_j) + \log \mu_{cj}(s_j) \right), \end{aligned}$$

where in the last step we replaced  $\theta_j$  with  $s_j$  for convenience. Here, we recognize  $\mu_{bj}$  and  $\mu_{cj}$  as the structured variational updates of Theorem 2. Identification of the fixed points can then be obtained by [57] (Corollary 2). For a rigorous discussion on convergence of the EM algorithm, we refer to [77] (Corollary 32), [24] (Chapter 6) and [57] (Section 3).  $\square$

Appendix D.16. Proof of Theorem 8

**Proof.** Substituting for  $q_a(s_a)$ , the node-local free energy becomes

$$\begin{aligned} F[q_a, f_a] &= \int q_a(s_a) \log \frac{q_a(s_a)}{f_a(s_a)} ds_a \\ &= \int q_a(s_a) \log \frac{q_{j|a}(s_j | s_{a \setminus j})}{f_a(s_a)} ds_a + \int q_a(s_a) \log q_{a \setminus j}(s_{a \setminus j}) ds_a \\ &= \int q_{a \setminus j}(s_{a \setminus j}) q_{j|a}(s_j | s_{a \setminus j}) \log \frac{q_{j|a}(s_j | s_{a \setminus j})}{f_a(s_a)} ds_a + \int q_{a \setminus j}(s_{a \setminus j}) q_{j|a}(s_j | s_{a \setminus j}) \log q_{a \setminus j}(s_{a \setminus j}) ds_a \\ &= \int q_{a \setminus j}(s_{a \setminus j}) \left[ \int q_{j|a}(s_j | s_{a \setminus j}) \log \frac{q_{j|a}(s_j | s_{a \setminus j})}{f_a(s_a)} ds_j \right] ds_{a \setminus j} + \int q_{a \setminus j}(s_{a \setminus j}) \log q_{a \setminus j}(s_{a \setminus j}) ds_{a \setminus j} \\ &= \mathbb{E}_{q_{a \setminus j}} \left[ D[q_{j|a} \| f_a] \right] - H[q_{a \setminus j}], \end{aligned}$$

where the first term expresses an expected Kullback–Leibler divergence, and the second term is a negative entropy. The only possibility for the local free energy to become finite, is when  $q_{j|a}(s_j|\mathbf{s}_{a\setminus j}) = f_a(\mathbf{s}_a) = \delta(s_j - g_a(\mathbf{s}_{a\setminus j}))$ . We then have:

$$F[q_a, f_a] = \begin{cases} -H[q_{a\setminus j}] & \text{if } q_{j|a}(s_j|\mathbf{s}_{a\setminus j}) = \delta(s_j - g_a(\mathbf{s}_{a\setminus j})) \\ \infty & \text{otherwise.} \end{cases}$$

□

#### Appendix D.17. Proof of Theorem 9

**Proof.** The proof is similar to Appendix D.16. Substituting for  $q_a(\mathbf{s}_a)$ , the node-local free energy becomes

$$\begin{aligned} F[q_a, f_a] &= \int q_a(\mathbf{s}_a) \log \frac{q_a(\mathbf{s}_a)}{f_a(\mathbf{s}_a)} \mathbf{d}\mathbf{s}_a \\ &= \int q_a(s_i, s_j, s_k) \log \frac{q_{ik|j}(s_i, s_k|s_j)}{f_a(s_i, s_j, s_k)} \mathbf{d}s_i \mathbf{d}s_j \mathbf{d}s_k + \int q_j(s_j) \log q_j(s_j) \mathbf{d}s_j \\ &= \mathbb{E}_{q_j} \left[ D[q_{ik|j} \| f_a] \right] - H[q_j]. \end{aligned}$$

In contrast to Appendix D.16, here we have a joint belief within the divergence with a single conditioning variable. Conditioning on  $s_j$  (or by symmetry  $s_i$  or  $s_k$ ) determines the realization of the other variables. Therefore, we have:

$$F[q_a, f_a] = \begin{cases} -H[q_j] & \text{if } q_{ik|j}(s_i, s_k|s_j) = \delta(s_j - s_i) \delta(s_j - s_k) \\ \infty & \text{otherwise.} \end{cases}$$

□

## References

- Blei, D.M. Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annu. Rev. Stat. Appl.* **2014**, *1*, 203–232. [CrossRef]
- Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
- Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
- Forney, G. Codes on graphs: Normal realizations. *IEEE Trans. Inf. Theory* **2001**, *47*, 520–548. [CrossRef]
- Loeliger, H.A. An introduction to factor graphs. *IEEE Signal Process. Mag.* **2004**, *21*, 28–41.
- Winn, J.; Bishop, C.M. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
- Yedidia, J.S.; Freeman, W.T.; Weiss, Y. *Understanding Belief Propagation and Its Generalizations*; Mitsubishi Electric Research Laboratories, Inc.: Cambridge, MA, USA, 2001.
- Cox, M.; van de Laar, T.; de Vries, B. A factor graph approach to automated design of Bayesian signal processing algorithms. *Int. J. Approx. Reason.* **2019**, *104*, 185–204. [CrossRef]
- Yedidia, J.S. An Idiosyncratic Journey beyond Mean Field Theory. In *Advanced Mean Field Methods*; The MIT Press: Cambridge, MA, USA, 2000; pp. 37–49.
- Yedidia, J.S.; Freeman, W.T.; Weiss, Y. *Bethe Free Energy, Kikuchi Approximations, and Belief Propagation Algorithms*; Mitsubishi Electric Research Laboratories, Inc.: Cambridge, MA, USA, 2001; p. 24.
- Dauwels, J. On Variational Message Passing on Factor Graphs. In Proceedings of the IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 2546–2550. [CrossRef]
- Zhang, D.; Wang, W.; Fettweis, G.; Gao, X. Unifying Message Passing Algorithms under the Framework of Constrained Bethe Free Energy Minimization. *arXiv* **2017**, arXiv:1703.10932.
- van de Laar, T.; Şenöz, I.; Özçelikkale, A.; Wymeersch, H. Chance-Constrained Active Inference. *arXiv* **2021**, arXiv:2102.08792.
- Smola, A.J.; Vishwanathan, S.V.N.; Eskin, E. Laplace propagation. In *NIPS*; The MIT Press: Cambridge, MA, USA, 2004; pp. 441–448.
- Minka, T. *Divergence Measures and Message Passing*. Available online: <https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/> (accessed on 24 June 2021).
- Yedidia, J.S. Generalized Belief Propagation and Free Energy Minimization. Available online: <http://cba.mit.edu/events/03.11.ASE/docs/Yedidia.pdf> (accessed on 24 June 2021).

17. Yedidia, J.S.; Freeman, W.; Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* **2005**, *51*, 2282–2312. [[CrossRef](#)]
18. Minka, T.P. Expectation Propagation for Approximate Bayesian Inference. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, 2–5 August 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 362–369.
19. Heskes, T. Stable fixed points of loopy belief propagation are local minima of the bethe free energy. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2003; pp. 359–366.
20. Kschischang, F.R.; Frey, B.J.; Loeliger, H.A. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory* **2001**, *47*, 498–519.
21. Hoffman, M.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic Variational Inference. *arXiv* **2012**, arXiv:1206.7051.
22. Archer, E.; Park, I.M.; Buesing, L.; Cunningham, J.; Paninski, L. Black box variational inference for state space models. *arXiv* **2015**, arXiv:1511.07367.
23. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1988.
24. Wainwright, M.J.; Jordan, M.I. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends® Mach. Learn.* **2008**, *1*, 1–305. [[CrossRef](#)]
25. Chertkov, M.; Chernyak, V.Y. Loop Calculus in Statistical Physics and Information Science. *Phys. Rev. E* **2006**, *73*. [[CrossRef](#)]
26. Weller, A.; Tang, K.; Jebara, T.; Sontag, D.A. Understanding the Bethe approximation: When and how can it go wrong? In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, Quebec City, QC, Canada, 23–27 July 2014; pp. 868–877.
27. Sibel, J.C. Region-Based Approximation to Solve Inference in Loopy Factor Graphs: Decoding LDPC Codes by Generalized Belief Propagation. Available online: <https://hal.archives-ouvertes.fr/tel-00905668> (accessed on 24 June 2021).
28. Minka, T. *From Hidden Markov Models to Linear Dynamical Systems*; Technical Report 531; Vision and Modeling Group, Media Lab, MIT: Cambridge, MA, USA, 1999.
29. Loeliger, H.A.; Dauwels, J.; Hu, J.; Korl, S.; Ping, L.; Kschischang, F.R. The Factor Graph Approach to Model-Based Signal Processing. *Proc. IEEE* **2007**, *95*, 1295–1322. [[CrossRef](#)]
30. Loeliger, H.A.; Bolliger, L.; Reller, C.; Korl, S. Localizing, forgetting, and likelihood filtering in state-space models. In Proceedings of the 2009 Information Theory and Applications Workshop, La Jolla, CA, USA, 8–13 February 2009; pp. 184–186. [[CrossRef](#)]
31. Korl, S. A Factor Graph Approach to Signal Modelling, System Identification and Filtering. Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich, Switzerland, 2005.
32. Pearl, J. Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. In Proceedings of the Second AAAI Conference on Artificial Intelligence, Pittsburgh, PA, USA, 18–20 August 1982; AAAI Press: Pittsburgh, PA, USA, 1982; pp. 133–136.
33. Heskes, T. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *J. Artif. Intell. Res.* **2006**, *26*, 153–190.
34. Särkkä, S. *Bayesian Filtering and Smoothing*; Cambridge University Press: London, UK; New York, NY, USA, 2013.
35. Khan, M.E.; Lin, W. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. *arXiv* **2017**, arXiv:1703.04265.
36. Logan, B.; Moreno, P. Factorial HMMs for acoustic modeling. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, USA, 15 May 1998; Volume 2, pp. 813–816. [[CrossRef](#)]
37. Hoffman, M.D.; Blei, D.M. Structured Stochastic Variational Inference. *arXiv* **2014**, arXiv:1404.4114.
38. Singh, R.; Ling, J.; Doshi-Velez, F. Structured Variational Autoencoders for the Beta-Bernoulli Process. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; p. 9.
39. Bamler, R.; Mandt, S. Structured Black Box Variational Inference for Latent Time Series Models. *arXiv* **2017**, arXiv:1707.01069.
40. Zhang, C.; Yuan, Z.; Wang, Z.; Guo, Q. Low Complexity Sparse Bayesian Learning Using Combined BP and MF with a Stretched Factor Graph. *Signal Process.* **2017**, *131*, 344–349. [[CrossRef](#)]
41. Wand, M.P. Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing. *J. Am. Stat. Assoc.* **2017**, *112*, 137–168. [[CrossRef](#)]
42. Caticha, A. Entropic Inference and the Foundations of Physics. In Proceedings of the 11th Brazilian Meeting on Bayesian Statistics, Amparo, Brazil, 18–22 March 2012.
43. Pearl, J. A Probabilistic Calculus of Actions. Available online: <https://arxiv.org/ftp/arxiv/papers/1302/1302.6835.pdf> (accessed on 24 June 2021).
44. Zoeter, O.; Heskes, T. Gaussian Quadrature Based Expectation Propagation. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Bridgetown, Barbados, 6–8 January 2005; p. 9.
45. Arasaratnam, I.; Haykin, S. Cubature Kalman Filters. *IEEE Trans. Autom. Control* **2009**, *54*, 1254–1269. [[CrossRef](#)]
46. Sarkka, S. Bayesian Estimation of Time-Varying Systems: Discrete-Time Systems. Available online: [https://users.aalto.fi/~ssarkka/course\\_k2011/pdf/course\\_booklet\\_2011.pdf](https://users.aalto.fi/~ssarkka/course_k2011/pdf/course_booklet_2011.pdf) (accessed on 24 June 2021).
47. Gelman, A.; Vehtari, A.; Jylänki, P.; Robert, C.; Chopin, N.; Cunningham, J.P. Expectation propagation as a way of life. *arXiv* **2014**, arXiv:1412.4869.



48. Deisenroth, M.P.; Mohamed, S. Expectation Propagation in Gaussian Process Dynamical Systems: Extended Version. *arXiv* **2012**, arXiv:1207.2940.
49. Teh, Y.W.; Hasenclever, L.; Lienart, T.; Vollmer, S.; Webb, S.; Lakshminarayanan, B.; Blundell, C. Distributed Bayesian Learning with Stochastic Natural-gradient Expectation Propagation and the Posterior Server. *arXiv* **2015**, arXiv:1512.09327.
50. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
51. Cox, M. Robust Expectation Propagation in Factor Graphs Involving Both Continuous and Binary Variables. In Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; p. 5.
52. Minka, T.; Winn, J.; Guiver, J.; Webster, S.; Zaykov, Y.; Yangel, B.; Spengler, A.; Bronskill, J. Infer.NET 2.6. 2014. Available online: <http://research.microsoft.com/infernet> (accessed on 23 June 2021).
53. Friston, K.J.; Harrison, L.; Penny, W. Dynamic causal modelling. *Neuroimage* **2003**, *19*, 1273–1302.
54. Mathys, C.D.; Daunizeau, J.; Friston, K.J.; Klaas, S.E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **2011**, *5*. [[CrossRef](#)]
55. Friston, K.; Kilner, J.; Harrison, L. A free energy principle for the brain. *J. Physiol.* **2006**, *100*, 70–87. [[CrossRef](#)]
56. Friston, K. The free-energy principle: A rough guide to the brain? *Trends Cogn. Sci.* **2009**, *13*, 293–301. [[CrossRef](#)]
57. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–38.
58. Dauwels, J.; Eckford, A.; Korl, S.; Loeliger, H.A. Expectation maximization as message passing—Part I: Principles and gaussian messages. *arXiv* **2009**, arXiv:0910.2832.
59. Bouvrie, P.; Angulo, J.; Dehesa, J. Entropy and complexity analysis of Dirac-delta-like quantum potentials. *Phys. A Stat. Mech. Appl.* **2011**, *390*, 2215–2228. [[CrossRef](#)]
60. Dauwels, J.; Korl, S.; Loeliger, H.A. Expectation maximization as message passing. In Proceedings of the International Symposium on Information Theory 2005, (ISIT 2005), Adelaide, Australia, 4–9 September 2005; pp. 583–586. [[CrossRef](#)]
61. Cox, M.; van de Laar, T.; de Vries, B. ForneyLab.jl: Fast and flexible automated inference through message passing in Julia. In Proceedings of the International Conference on Probabilistic Programming, Boston, MA, USA, 4–6 October 2018.
62. Bezanson, J.; Edelman, A.; Karpinski, S.; Shah, V. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* **2017**, *59*, 65–98. [[CrossRef](#)]
63. Şenöz, I.; de Vries, B. Online Variational Message Passing in the Hierarchical Gaussian Filter. In Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, Denmark, 17–20 September 2018; pp. 1–6. [[CrossRef](#)]
64. Mathys, C.D. *Uncertainty, Precision, and Prediction Errors*; UCL Computational Psychiatry Course; UCL: London, UK, 2014.
65. Şenöz, I.; de Vries, B. Online Message Passing-based Inference in the Hierarchical Gaussian Filter. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 2676–2681. [[CrossRef](#)]
66. Podusenko, A.; Kouw, W.M.; de Vries, B. Online Variational Message Passing in Hierarchical Autoregressive Models. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 1337–1342. [[CrossRef](#)]
67. Welling, M. On the Choice of Regions for Generalized Belief Propagation. *arXiv* **2012**, arXiv:1207.4158.
68. Welling, M.; Minka, T.P.; Teh, Y.W. Structured Region Graphs: Morphing EP into GBP. *arXiv* **2012**, arXiv:1207.1426.
69. Loeliger, H.A. Factor Graphs and Message Passing Algorithms—Part 1: Introduction, 2007. Available online: [http://www.crm.sns.it/media/course/1524/Loeliger\\_A.pdf](http://www.crm.sns.it/media/course/1524/Loeliger_A.pdf) (accessed on 3 April 2019).
70. Caticha, A. Relative Entropy and Inductive Inference. *AIP Conf. Proc.* **2004**, *707*, 75–96. [[CrossRef](#)]
71. Ortega, P.A.; Braun, D.A. A Minimum Relative Entropy Principle for Learning and Acting. *J. Artif. Intell. Res.* **2010**, *38*, 475–511.
72. Shore, J.; Johnson, R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37. [[CrossRef](#)]
73. Engel, E.; Dreizler, R.M. *Density Functional Theory: An Advanced Course*; Theoretical and Mathematical Physics; Springer: Berlin/Heidelberg, Germany, 2011. [[CrossRef](#)]
74. Boyd, S.P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2004.
75. Lanczos, C. *The Variational Principles of Mechanics*; Courier Corporation: North Chelmsford, MA, USA, 2012.
76. Ahn, S.; Chertkov, M.; Shin, J. Gauging Variational Inference. Available online: <https://dl.acm.org/doi/10.5555/3294996.3295048> (accessed on 24 June 2021).
77. Tran, V.H. Copula Variational Bayes inference via information geometry. *arXiv* **2018**, arXiv:1803.10998.